# Predicting COVID-19 with Time Series Forecasting using Machine Learning

**Sujay S**

*Student, Department of Electronics and Communication Engineering, Velammal Engineering College, Chennai, Tamil Nadu, India*

---***---

**Abstract -** *Time Series analysis is being widely used in the fields of corporate sectors, medical sciences, weather forecasting, and stock prices to predict future values based on previously observed values. In this research, Time Series data of COVID-19 (SARS-CoV-2) obtained from a dataset obtained from worldometers has been to predict the prognosis of the virus in the forthcoming days. The dataset contains summed up values of the daily new cases, recovered cases, and deaths from 187 countries right from the day from which the first case was reported. After exploring and splitting the data into train and test sets to ensure the fidelity of the model, a Machine Learning model called Prophet which was introduced by Facebook has been used to train and test on the observed data. Finally, it has been used to draw probabilistic insights for the future on the new cases, recovered cases, and new deaths for the next month.*

***Key Words***: Machine Learning, Time Series Forecasting, Facebook Prophet, Exploratory Data Analysis

## 1. INTRODUCTION

Coronavirus disease (COVID-19) is a respiratory disease caused by a newly discovered coronavirus [1]. The common symptoms of this disease are fever, dry cough, and tiredness. Other symptoms are aches and pains, nasal congestion, headache, sore throat, loss of taste and smell called Anosmia. It also causes a condition called Parosmia where the scents that one used to find pleasant may now become unbearable. One cause of parosmia symptoms is olfactory damage due to the virus. COVID-19 can damage the lungs, causing pneumonia.

The virus can exacerbate through the respiratory tract and enter into a person's lungs. This causes damage to the air sacs or alveoli, that can fill with fluid. This progression then constraints a person's ability to take in oxygen. Continuous oxygen deprivation can damage many of the body's organs, causing kidney failure, heart attacks, and other life-threatening conditions.

People who have pre-existing conditions such as cancer, diabetes, high blood pressure, kidney or liver disease, including but not limited to asthma are at most risk of COVID-19 pneumonia. People over the age of 65 years are more prone to the intense effects of this disease. The disease has turned into a widespread pandemic where the cases and deaths seem to surge rapidly day by day.

This research intends to uncover the prognosis of various parameters involved with this virus such as the increase of new cases, recoveries and deaths daily worldwide with the help of a machine learning technique called Prophet [3] model which was developed and introduced by Facebook.

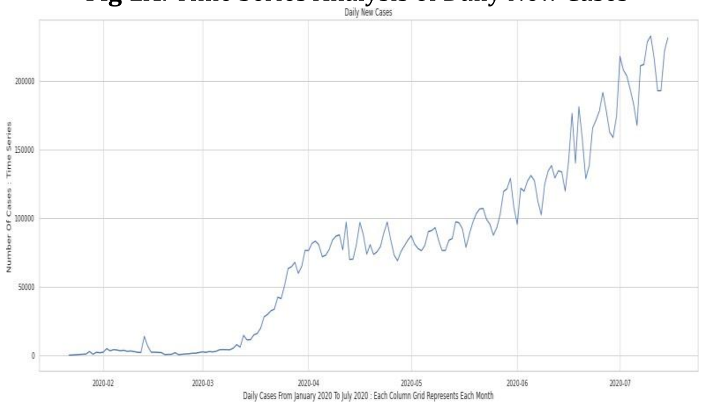
## 2. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is the deciding factor for selecting the model for the data. First, let us understand the impact of the novel coronavirus in this world through graphical visualizations [4].

Before plotting the data, it has been pre-processed by converting the date column present in it to a date-time format using a function that pre-exists in Python. After that, the data has been plotted using a Python library called matplotlib [5] and the forthcoming graphs are represented. Fig 1A shows the Time Series analysis of the daily new cases right from the day from which the first case was reported till the date 2020-07-15. The x-axis shows the year and month in date-time format respectively.

Fig 1B shows the Time Series analysis [6] of daily recovered cases and the Fig 1C shows the Time Series Analysis of daily deaths.

**Fig 1A**: Time Series Analysis of Daily New Cases



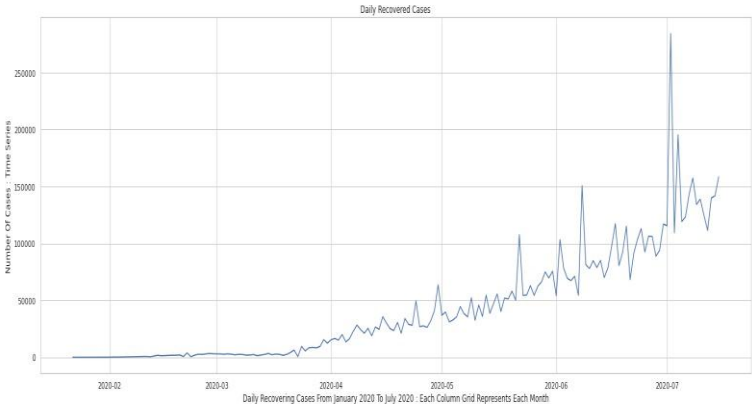


**Fig 1B**: Time Series Analysis of Daily Recovered Cases

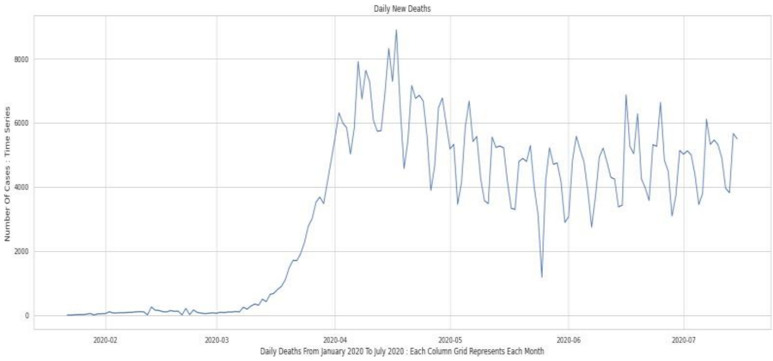


**Fig 1C**: Time Series Analysis of Daily New Deaths

Every Time Series analysis [6] can be categorized into the components namely Trend, Seasonality and Cyclic variations. The trend shows the general tendency of the data to increase or decrease during a long period. It can either be linear or nonlinear. Seasonal variations are the rhythmic forces which operate regularly and periodically over a span of less than a year. Cyclic variations operate themselves over a span of more than one year are the cyclic variations. Considering the plots respective to this data, it can be concluded that Fig 1A, 1B and 1C show a nonlinear variation which can be put into the category called Trend.

## 3. DATA PREPROCESSING & MODEL SELECTION

To improve the model's efficiency, the day from which there is a significant increase in the values are only included for training the model where it starts to show a linear trend. The following plots Fig 2A, 2B and 2C are plotted after dropping the redundant values using pandas [7] in the data. It is that from 2020-04-15 the cases seem to surge for all of the parameters.

ARIMA models can lead to extremely long fitting times and require modelling expertise that many forecasting novices would not have. The prophet [3] is a procedure for forecasting time series data and is robust to missing data and shifts in the trend, and typically handles outliers well. Prophet library utilizes the additive regression model y(t) comprising the following components:

$$y(t)=g(t)+s(t)+h(t)+\epsilon t$$

Where Trend g(t) represents models non-periodic changes, Seasonality s(t) represents periodic changes and Holidays component h(t) contributes information about holidays and events [3]

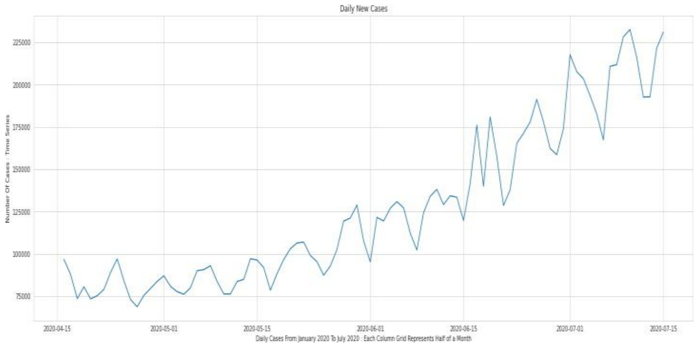


**Fig 2A**: Processed Time Series Analysis of Daily New Cases

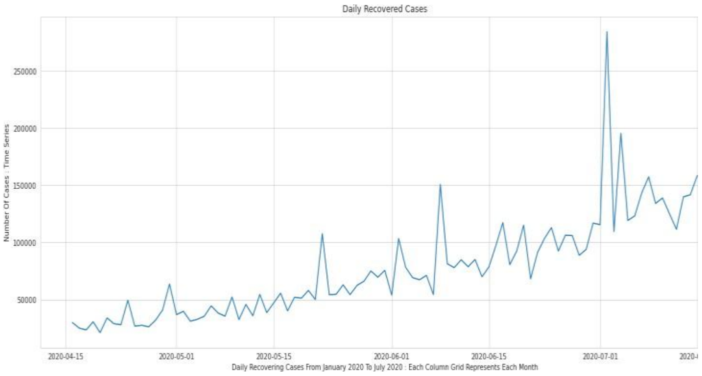


**Fig 2B**: Processed Time Series Analysis of Daily Recovered Cases

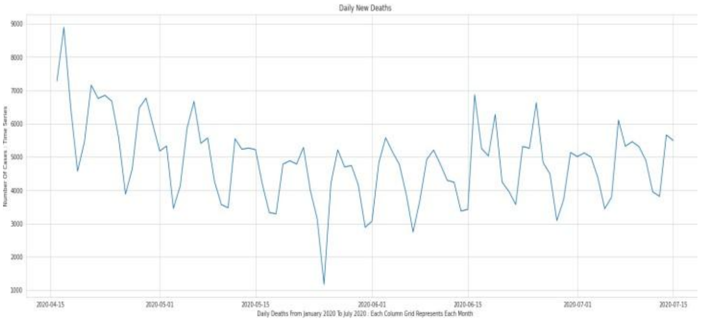


**Fig 2C**: Processed Time Series Analysis of Daily New Deaths

## 4. MODEL TRAINING & TESTING

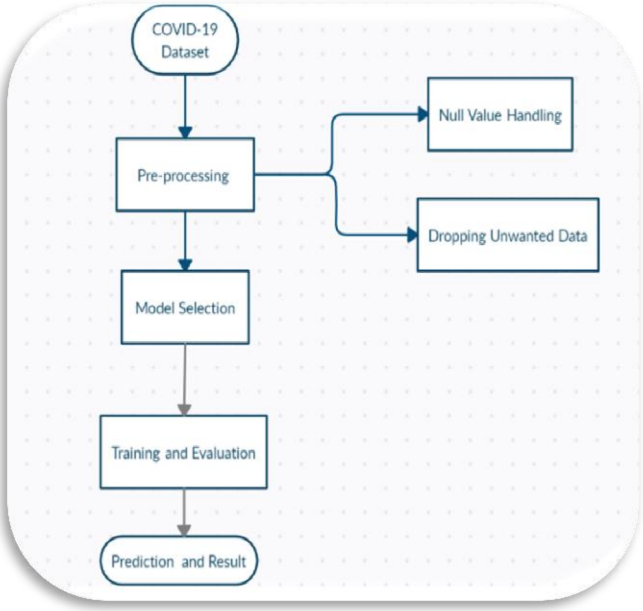
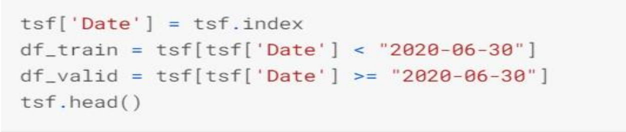


**Fig 3**: Model Architecture

Fig 3 represents the complete workflow of the project. The data is now split into train and test sets called df_train and df_valid respectively. Dates before 2020-06-30 have been used to train the model and which occur between 2020-06- 30 and 2020-07- 15 has been used to test the model's accuracy.



```
tsf['Date'] = tsf.index
df_train = tsf[tsf['Date'] < "2020-06-30"]
df_valid = tsf[tsf['Date'] >= "2020-06-30"]
tsf.head()
```

**Fig 4**: Code for Splitting Data into Train and Test Set

Fig 4 shows the snippet used to split the data into training and validation sets. In Fig 4A, 4B and 4C the orange line represents the data on which the model has trained, the blue line represents the observed values on the validation set and the green line represents the values predicted by the Prophet model. The model has performed well in all the cases especially in predicting the new recovered cases and the new deaths.

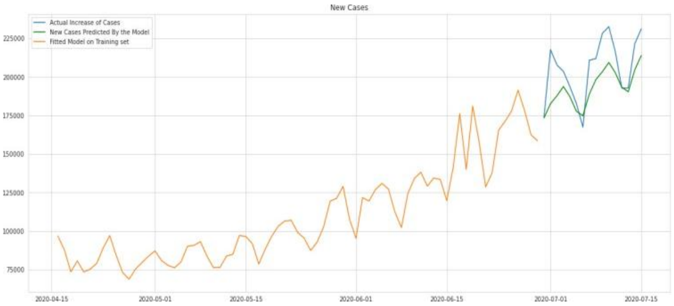


**Fig 4A:** Time Series Forecasting of Daily New Cases

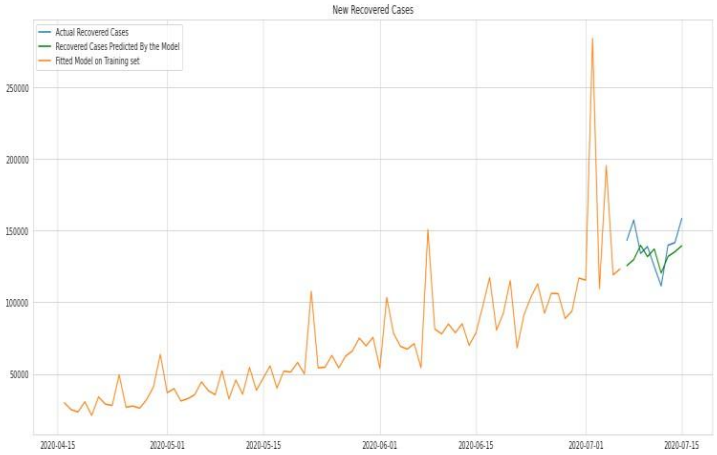


**Fig 4B**: Time Series Forecasting of Daily Recovered Cases by Prophet

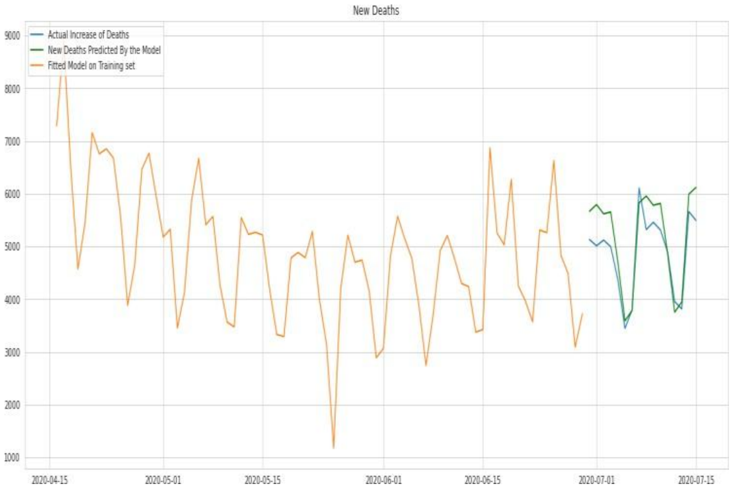


**Fig 4C**: Time Series Forecasting of Daily Deaths by Prophet

A new dataset has been created containing the future dates for the next 30 days and the values for which the model predicts remain null so that the model can predict it.

Now the model is allowed to predict the daily new cases, recovered cases and new deaths between 2020-07-07 and 2020-08- 15.

The Fig 5A, 5B and 5C show the future forecasts for daily new cases recovered cases and deaths for the forthcoming month. On explicitly stating, the model is predicting the dates between 2020- 07-07 and 2020-08-15. The blue line in the plots shows the values on which the model trained and the green line shows the future forecast prediction by the model.

## 5. FORECASTING THE FUTURE

A new dataset has been created containing the future dates for the next 30 days and the values for which the model predicts remain null so that the model can predict it.
Now the model is allowed to predict the daily new cases, recovered cases and new deaths between 2020-07-07 and 2020-08- 15.

The Fig 5A, 5B and 5C show the future forecasts for daily new cases recovered cases and deaths for the forthcoming month. On explicitly stating, the model is predicting the dates between 2020- 07-07 and 2020-08-15. The blue line in the plots shows the values on which the model trained and the green line shows the future forecast prediction by the model.

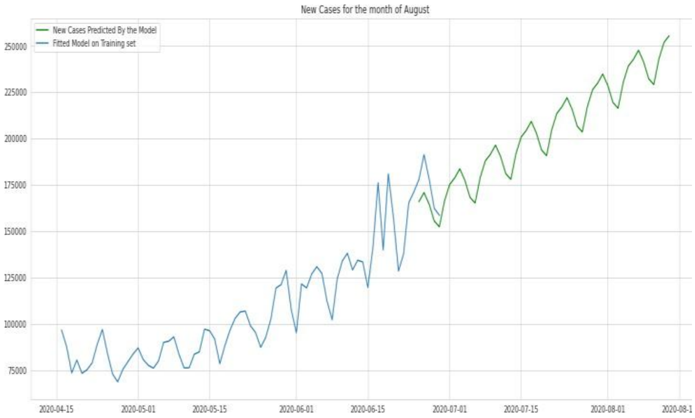


**Fig 5A**: Future Forecasting of Daily New Cases by Prophet

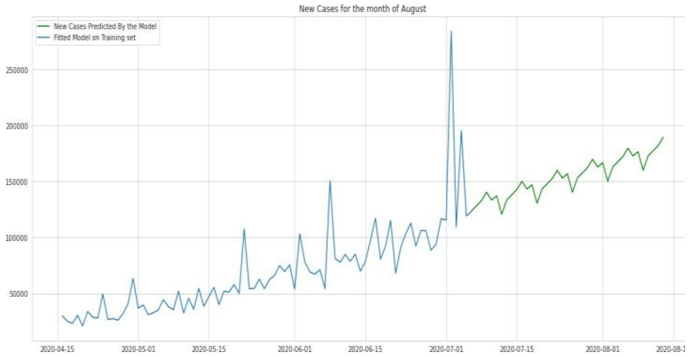


**Fig 5B**: Future Forecasting of Daily New Recovered Cases

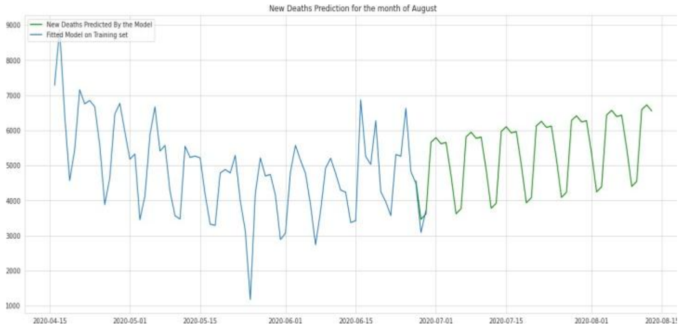


**Fig 5C**: Future Forecasting of Daily New Deaths

## 6. CONCLUSION

The model has predicted that there will be an increase in the daily new cases, recovered cases and the new deaths for the next month.

In this research, the significance of the prophet model and the prognosis of the deadly COVID-19 has been represented. Using the cumulative data of 187 countries, statistical analysis was performed to gather insights into the data. Finally, using Machine Learning techniques, a future forecast of the Time Series data was predicted and represented through the same graphs.

## REFERENCES

[1] World Health Organization. (2020). Coronavirus disease 2019 (COVID-19): situation report, 30. World Health Organization.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. CoronaTracker: Worldwide COVID-19 Outbreak Data Analysis and Prediction. Bull World Health Organ. K. Elissa, "Title of paper if known," unpublished.

[3] Taylor SJ, Letham B. 2017. Forecasting at scale. PeerJ Preprints 5:e3190v2.

[4] An interactive web-based dashboard to track COVID-19 in real-time. February 19, 2020 https://doi.org/10.1016/ S1473- 3099(20)30120-1.

[5] Matplotlib: A 2D Graphics Environment, J.D. Hunter et al., Computing in Science & Engineering, IEEE Computer SOC.

[6] Whiteley, P. Time series analysis. Qual Quant 14, 225–247 (1980).

[7] Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51- 56 (2010).

[8] https://www.kaggle.com/imdevskp/corona-virus-report ( Last accessed on July 20 )

## BIOGRAPHIES

**Sujay S** Student at Velammal Engineering College, Chennai, pursuing B.E., ECE department III year. High school in T.I. Matriculation Higher Secondary School, Ambattur, Chennai. Currently learning latest Machine Learning Algorithms in solving real-world problems and implementing extensive research in the field of predictive modelling, Exploratory Data Analysis, model deployment & Containerization.