Covid -19 Prediction and Analysis

Shubhang Shukla[#], Sarthak Mishra[*], Dr Ruchika Malhotra[#]

[#]*Software Department, Delhi Technological University*

Bawana Rd, Delhi Technological University, Shahbad Daulatpur Village, Rohini, New Delhi, Delhi 110042

[1]shubhangnshukla_2k19me239@dtu.ac.in

[3]sarthakmishra_2k19pe058@dtu.ac.in

***Abstract*— Covid-19 pandemic has affected the lives of millions of people and crippled the health systems of many countries. A significant factor was its speed of spread and unpredictability. In our project we have tried to analyze the data of various states across India to understand which were the most affected and also use Forecasting techniques such as time series to predict the cases for the next week. The predictions were not as accurate as expected due to the large number of factors that need to be considered.**

***Keywords*— Covid-19, Predictive Analytics, Time-Series Forecasting, Arima Model, Prophet Model, Holt's Model**

I. Introduction

The WHO declared COVID-19 as a pandemic in February 2020. What started in Wuhan in China in a matter of 30 days spread to more than half of the countries. As the entire world grappled with this new virus, wearing masks and social distancing was the only way out. Almost a year later as vaccines were in the last stage of development and cases seemed to decline we were hit by the 2nd wave.

This unpredictability in infections and waves in which it is coming points towards the need to be able to forecast the cases in order to be better prepared. Using Machine Learning techniques some amount of prediction can be made to understand the trends.

II. Dataset Collection

The data which is collected through surveys, records, etc is called raw data. This data may be structured or unstructured, small or big, simple or complex. Large amounts of Structured data are stored in what is called as Data Warehouses while large amounts of unstructured data is stored in Data Lake.
Data for COVID-19 is gathered from hospitals, testing labs and other primary and secondary healthcare centers etc and is collected at the district, municipality, state and national level. At the world level bodies such as WHO and John Hopkins University are maintaining world wide databases which are updated everyday.

There are various governmental and non governmental bodies which are maintaining Covid 19 Databases. In India both the central and state governments have been maintaining detailed records of cases, trends and vaccinations. The Ministry of Health and Family Welfare is in charge of the COWIN dashboard and has been providing a range of facilities for tracking, services and analytics. Link

III. data pre-processing

A. Feature Selection

The dataset which we took was huge with many attributes. Keeping only relevant columns out of the large database from our analysis point of view was the first step before starting the analysis.The file in csv format, was first studied, verified and cleaned in Excel and then loaded into Python for analysis.

B. Data Structuring

The data sourced directly contained everything bundled together in a single data frame hence structuring the data was done as required for the relevant analysis.

C. Data Cleaning

The data first had to be cleaned by removing the Null and missing values. There were a lot of missing values which means that for a particular date, the data for covid cases was not updated. Datatype also had to be changed such as the time was in a string format which was converted to datetime.

D. Outlier Detection and Data Scaling

The dataset also contained cumulative or total values of confirmed cases/ deaths etc in between the rows. These are outliers and they had to be made into a separate data frame. The values of total cases and deaths etc number in millions, hence for the purpose of understanding, comparing and plotting we had to scale them by a factor of E-6. Further for calculating the percentages we had to normalize the values.

IV. exploratory analysis

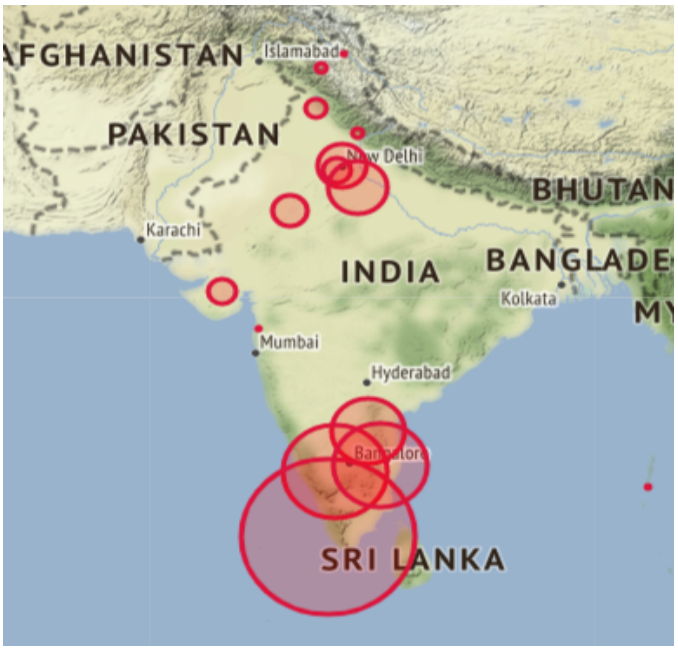
Exploratory Data Analysis was carried out after the data was cleaned and structured. EDA helps in getting familiar with the data and understanding basic trends and summary statistics. We divided our EDA approach into three parts. The first part was an analysis of the states followed by the entire country and finally for the cumulative of all countries. The trends of rise in cases, active cases and deaths were analyzed and compared across the states.

We also compared the infections in India with the total cases in the world as a progression with time. Countries such as Italy , US, China were also plotted to analyse the cases.

New metrics such as mortality rate were calculated for analysis.Data Visualization is an important part of EDA and several graphs were plotted to compare and contrast different metrics across geography.

DATA VISUALIZATION

COVID CASES ON MAP OF INDIA



STATE WISE CONFIRMED AND RECOVERED CASES

Fig. 1 The map shows the maximum number of cases overall were in South India but in the second wave North India was the worst affected and saw a large rise in infections.

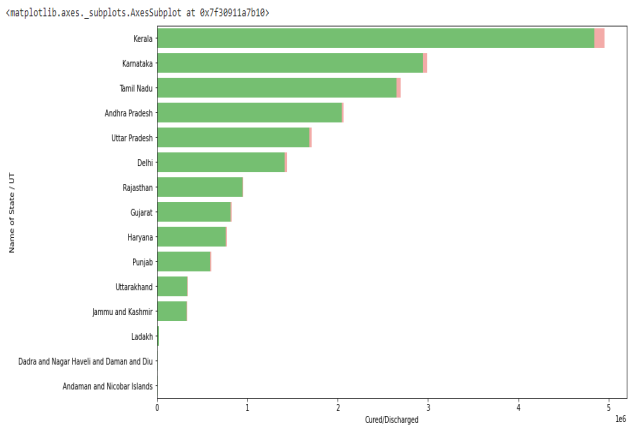


Fig. 2 Recovery vs Total Confirmed Cases

Fig. 2 shows the confirmed vs recovered cases for the top 10 affected states. A positive sign is that recovery rates were very high and 96 % of people recovered on an average. While Kerala had the highest number of cases in the country it also had the highest recovery rates.

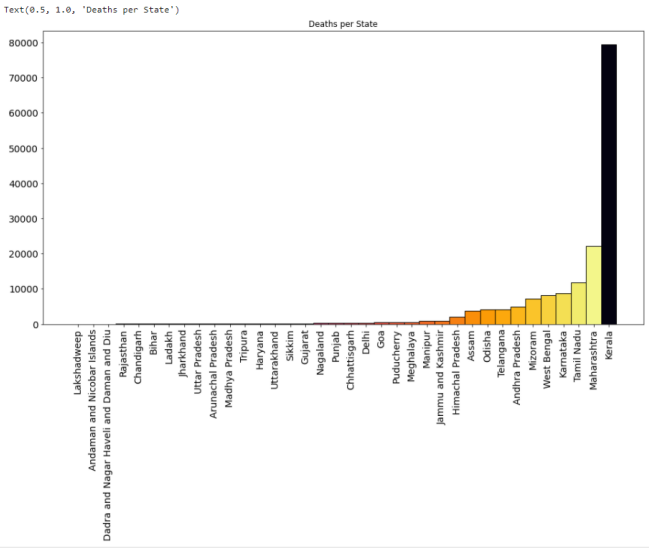
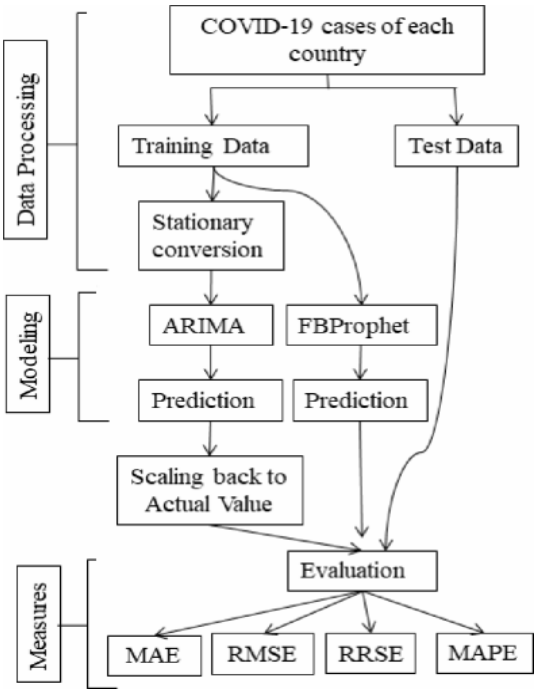


Fig. 3 Number of Deaths State Wise in ascending order

V. TIME SERIES FORECASTING

We used the following methodology to carry out our analysis. For the analysis, we have split the datasets of confirmed, active, recovered, and death cases into training and testing. We performed prediction after removing trends wherever applicable, and used statistical measures to evaluate the performance.



Timer series are a set of observations taken at different times/intervals. The model uses timestamped historical data to

predict the future values. These are extensively used in statistics, econometrics, mathematical finance, weather forecasting etc. Time series are usually broken down into different components such as trend, seasonality, cyclicity, noise etc.

E. Prophet Model

This model works best with time series that have strong seasonality and several seasons of historical data. Prophet can be considered a nonlinear regression model of the form.

$$yt=g(t)+s(t)+h(t)+\varepsilon t$$

where g(t) describes a piecewise-linear trend (or "growth term"), s(t) describes the various seasonal patterns, h(t) captures the holiday effects, and εt is a white noise error term.

The knots (or changepoints) for the piecewise-linear trend are automatically selected if not explicitly specified. Optionally, a logistic function can be used to set an upper bound on the trend.The seasonal component consists of Fourier terms of the relevant periods. By default, order 10 is used for annual seasonality and order 3 is used for weekly seasonality.Holiday effects are added as simple dummy variables.The model is estimated using a Bayesian approach to allow for automatic selection of the changepoints and other model characteristics

F. Arima Model

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q, where: p is the order of the AR term, q is the order of the MA term, d is the number of differencing required to make the time series stationary. If a time series has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'. The equation is as follows:

$$z_t = \alpha + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + w_t$$

G. Holt's Model

Holt's two-parameter model, also known as linear exponential smoothing, is a popular smoothing model for forecasting data with trend. Holt's model has three separate equations that work together to generate a final forecast. The first is a basic smoothing equation that directly adjusts the last smoothed value for last period's trend. The trend itself is updated over time through the second equation, where the trend is expressed as the difference between the last two smoothed values. Finally, the third equation is used to generate the final forecast. Holt's model uses two parameters, one for the overall smoothing and the other for the trend smoothing equation. The method is also called double exponential smoothing or trend-enhanced exponential smoothing.

VI. RESULTS

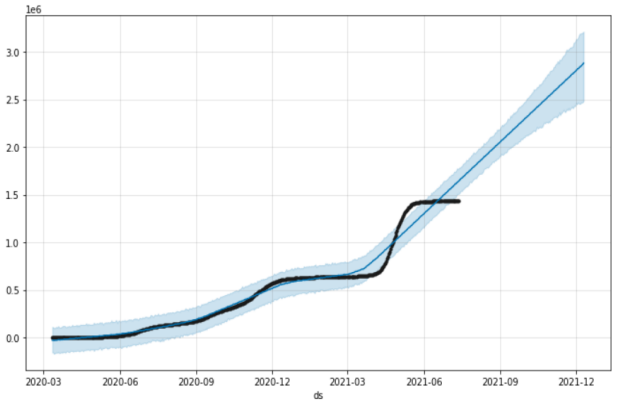


Fig. 4 Prediction based on Prophet Model for future cases

The Prophet model predicts a rise in the cases in the future. This could mean the onset of a third wave or the prediction may be completely off if through sufficient means we are able to flatten the COVID 19 curve.

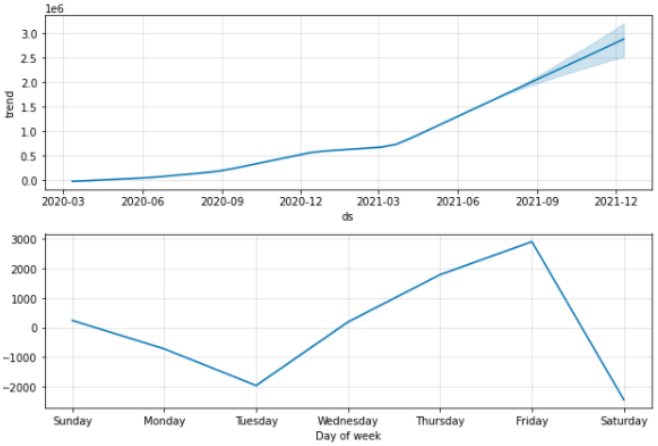


Fig. 5 Prediction of trend and weekly trends by the Prophet Model

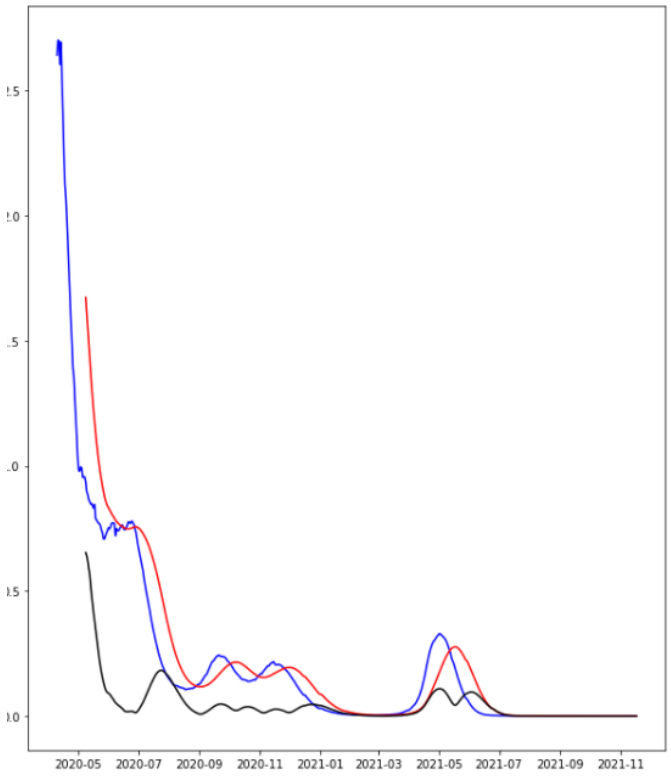


Fig. 5 Prediction on Log scale of the ARIMA Model

The Log transformation was carried out to make the series stationary and the values for the ARIMA model were chosen based on ACF and PACF graphs.

COMPARING THE THREE MODELS BASED ON EVALUATION METRICS

METRIC	Time Series Forecasting Model		
	ARIMA	PROPHET	HOLT
MAE	437312.63	739336.250	517612.64
RMSE	488652.73	794901.778	592111.12

As we can see the predictions are very accurate as was expected due to underfitting and bias in the time series models. However in the scope of things ARIMA model performed the best with the least error of the other two.

VII. Conclusions

To predict pandemics like COVID-19 just a time series of the data is not sufficient to generate accurate models. There are several factors that have to be considered in order to actually forecast the cases. In particular, factors like the region, population size, demographics, vaccination rate, government norms and regulations, containment measures etc have to be considered.

Further in this project we solely used the time series models however a variety of other Machine Learning and Deep Learning Algorithms can be applied to get better predictions and consider more parameters.

Acknowledgment

We would like to thank Dr Ruchika Malhotra for her guidance and support through our project. This project would have not been possible without the knowledge of the subject for which we would also like to thank the countless authors for the literature in this subject. We are extremely grateful to our family for providing us the right environment and encouragement to experiment with new things and pursue what we want to.

References

[1] N. Kumar and S. Susan, "COVID-19 Pandemic Prediction using Time Series Forecasting Models," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225319.

[2] COVID-19 in India, https://www.kaggle.com/ sudalai rajkumar/covid19-inindia?select=AgeGroupDetails.csv, (Last Accessed 12.05.2020)

[3] Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, et al., "World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19)", *International Journal of Surgery*, 2020.

[4] Shaun S Wulff, "Time series analysis: Forecasting and control", *Journal of Quality Technology*, vol. 49, no. 4, pp. 418, 2017.

[5] Duccio Fanelli and Francesco Piazza, "Analysis and forecast of covid-19 spreading in China, Italy and France", *Chaos Solitons & Fractals*, vol. 134, pp. 109761, 2020.

[6] Christopher JL Murray et al., "Forecasting covid-19 impact on hospital bed-days icu-days ventilator-days and deaths by the US state in the next 4 months", *MedRxiv*, 2020.

[7] Nalini Chintalapudi, Gopi Battineni and Francesco Amenta, "Covid-19 disease outbreak forecasting of registered and recovered cases after sixty day lockdown in italy: A data driven model approach", *Journal of Microbiology Immunology and Infection*, 2020.

[8] S. Chordia and Y. Pawar, "Analyzing and Forecasting COVID-19 Outbreak in India," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 1059-1066, doi: 10.1109/Confluence51648.2021.9377115.