Title: Predicting the growth and trend of COVID-19 (Indian States) using Machine Learning

Abstract:
The highly infectious novel coronavirus disease (COVID-19) was first detected in Wuhan, China, on 31st December 2019 and eventually spread across 196 countries and territories worldwide, infecting millions of people. In India, a densely populated country with around 1.3 billion population, the disease was first reported on 30th January 2020 in a Wuhan returned student of Kerala. The cumulative positive cases are rapidly increasing with time. Many studies have been accomplished to track the disease, its growth prediction, and possible policies to manage. However, most of this prior research focused on the number of infected people in the entire country, which can sometimes be misleading considering India's vast diversity. Hence in this study, we focus on the number of infected people in each Indian state (with sufficient data for prediction) and predict the number of infected people for that state in the next 60-90 days. We hope that such state-wise predictions would help the respective state governments to utilize their healthcare resources and take proper measures properly. This information can enable the authorities to lift the lockdown in a phased manner.

Introduction:
Coronavirus or COVID 19 does not require much introduction as it has impacted our lives very badly. However, let us have some quick looks which will help us to understand this situation better. On December 31, 2019, China reported WHO about unusual pneumonia cases being witnessed in the city of Wuhan. It was the very beginning of COVID-19. Since then, the cumulative incidence of the causative virus (SARS-CoV-2) is rapidly increasing and has affected around 220 countries, with a death of 2.5 million people and still counting. These viruses can cause respiratory symptoms in humans, along with other symptoms of common cold and fever. There are no specific treatments for coronaviruses to date. However, one can avoid infection by maintaining basic personal hygiene and social distancing from infected persons. India had also made it to the list of victims of this virus. The first case was reported at the end of January 2019.

Around 11 million cases have been reported in India till now, with 157K deaths. Besides people's health, the coronavirus outbreak has also affected the economy of the country. One such outcome is depicted in the price hike of vehicle diving fuels as COVID-19 has affected imports and exports of the country. Also, the lockdown loss needs to be compensated. The online study was sought as one of the major revolutions of the 21st century. However, no one had expected it to come in handy in such a situation. Students have been away from their educational institutions for more than a year now.
After following an exponential growth, fortunately, cases in India are decreasing now. The exact reason behind it is hard to predict. Scientists are giving credit to the proper execution of rules and preventive measures, warm climatic conditions, pre-covid preparations. Also, the distribution of vaccines has played its part. Still, we cannot expect the conditions to become normal again. States like Maharashtra are again witnessing the increase in cases. Since the beginning of February, Amravati has recorded more than 10,000 cases and over 66 deaths from Covid. More than 1,000 were receiving treatment for the disease this week. The positivity rate is in frightful double digits. Amravati and a few other districts in Maharashtra have been again locked down. However, states like Uttarakhand, Himachal Pradesh have recorded relatively fewer cases. As of February 23, there were 220 active cases in the state of Himachal Pradesh.
The world requires efficient methodologies and research to confront this chaotic situation. Deep learning presents itself as one such tool today. For example, ML and AI are used to augment the diagnosis and screening process of the identified patient with radio imaging technology. The list of such examples is endless. Wrapping covid datasets with Machine Learning (ML) and Artificial Intelligence (AI), researchers can forecast where and when the disease is likely to spread and notify those regions to match the required arrangements. Our group makes a small endeavour in similar regard. We are trying to analyse and predict the pandemic situation in different states of India. We are using Machine learning algorithms like Levenberg-Marquardt (LM) for curve fitting. We will be trying and testing various distributions and try to present the best fit in different scenarios.

Literature Review:
COVID-19 is caused by SARS-Cov-2, a newly emergent coronavirus that was first recognized in Wuhan, China, in December 2019. SARS-CoV-2 is a new coronavirus closely related to SARS-CoV and genetically clusters within *Betacoronavirus* subgenus *Sarbecovirus*. The first whole-genome sequence was published on January 5, 2020, and thousands of genomes have been sequenced since this date. Over 57 000 genome sequences have been deposited in the GISAID EpiCoV database. A meta-analysis of different time estimates to the virus's last common ancestor indicates that the pandemic could have started between October 6 and December 11, 2019 [ii]. (SAGE, 2020)
There is a need for innovative solutions to develop, manage, and analyse big data on infected subjects' growing network, patient details, their community movements and integrate with clinical trials and pharmaceutical, genomics, and public health data. The Machine Learning (ML) and Data science community are striving hard to improve the forecasts of epidemiological models and analyse the information flowing over Twitter to develop management strategies and the assessment of the impact of policies to curb its spread. Various datasets in this regard have been openly released to the public. However, there is a need to capture, develop and analyse more data as the COVID-19 grows worldwide. [i] (Shreshth Tuli, 2020). The base paper analyses and predicts the growth of COVID – 19 worldwide; however, this paper focuses on using linear regression models for COVID -19 prediction on a state-level. These models have already been

used to predict epidemics like COVID-19 worldwide, including China, Ebola outbreak in Bomi, Liberia (2014). We have used Indian COVID-19 data available publicly.
There are a few works that are based explicitly on Indian COVID-19 data. Das [iii] has used the epidemiological model to estimate the primary reproduction number at national and some state levels. Ray et al. [iv] used a predictive model for case-counts in India. They also discussed hypothetical interventions with various intensities and provided projections over a time horizon. Both the articles have used the SIR (susceptible-infected-removed) model for their analysis and prediction. As we discussed earlier, considering the great diversity in every aspect of India and its vast population, it would be a much better idea to look at each of the states individually. The study of each of the states individually would help decide further actions to contain the disease's spread, which can be crucial for the specific states only. [v] (Ghosh, 2020). In this article, the author mainly focused on the SIS model and the logistic and the exponential models at each state (restricting to only those states with enough data for prediction). The SIS model considers the possibility that an infected individual can return to the susceptible class on recovery because the disease confers no long-standing immunity against reinfection. WHO is aware of these reports of patients who were first tested negative for COVID-19 using PCR (polymerase chain reaction) testing and then after some days tested positive again [vi]. (Ghosh, 2020).

Dataset Review:
The dataset that we will be using is available on the Internet and available for public use. The main dataset on which the research has been carried is from "Our World in Data" by Hannah Ritchiee. It can be found on https://ourworldindata. org/coronavirus-source-data. This dataset contains a lot of unique parameters like reproduction rate, information about hospital admissions.
But it is noteworthy that the datasets, which contain information about different Indian states, do not have as much information as we have for India. Therefore, one should not expect much robust prediction on the state's data. These datasets can be found on the given GitHub link. https://github.com/covid19india/api.
It should be noted that the dataset is biased to many factors due to the fact that different states like Maharashtra, Andhra Pradesh, Uttar Pradesh, and Tamil Nadu have imposed strict lockdowns, whereas states like Punjab, Haryana have a less restricted lockdown.

Description of some of the important variables is presented below:

Variable name	Description
date	Date of observation
total_cases	Total confirmed cases of COVID-19
new_cases	New confirmed cases of COVID-19
new_cases_smoothed	New confirmed cases of COVID-19 (7-days smoothed)
total_deaths	Total deaths attributed to COVID-19
new_deaths	New deaths attributed to COVID-19
new_deaths_smoothed	New deaths attributed to COVID-19 (7-days smoothed)
total_cases_per_million	Total confirmed cases of COVID-19 per 1,000,000 people

new_cases_per_million	New confirmed cases of COVID-19 per 1,000,000 people
total_deaths_per_million	Total deaths attributed to COVID-19 per 1,000,000 people
new_deaths_per_million	New deaths attributed to COVID-19 per 1,000,000 people
total_tests_per_thousand	Total tests for COVID-19 per 1,000 people
new_tests_per_thousand	New tests for COVID-19 per 1,000 people

Objectives and Hypotheses:
The main objective of our study is to explore the trend of covid 19 in Indian states. Most of the prior research has focused on India as a single unit. However, extending the research to the state-level is important as the State governments have played a separate role in the control of Covid. Moreover, analysis at a smaller level will give more accuracy in implementing measures according to need. Through this study, we will try to answer the following questions:

1. Which states are under adverse conditions and which are under good control?
2. What is the expected number of cases after the end of November?
3. Which state will get rid of COVID-19 first (i.e., Number of cases < 1)?
4. When will COVID-19 vanish in India?
5. Which state has a chance to suffer from another wave and when?
6. What is Mortality Rate (State-Wise)?

Answering these can help authorities to know the scenario for lifting up the lockdown in a phased manner, especially in states which fall under the green zone.

Methodologies and Methods:
COVID-19 predictions are fundamentally important for rationalizing decisions, planning, and mentality, but also challenging due to the innate uncertainty of the complex, dynamic and global COVID-19 pandemic as a typical wicked problem. Here, we used mathematical modelling to predict the trend of patient diagnosis in Indian states, with the aim of easing anxiety regarding the emergent situation. According to all diagnosis numbers from the WHO website and combining with the transmission mode of infectious diseases, the mathematical model was fitted to predict the future trends of outbreaks. The global trend was approximately exponential, with an increased rate of 10-fold every 19 days. After finding a suitable graph for our dataset, we need to find appropriate values for our parameters (α, β, γ) to minimize the error between the predicted cases $(y = f(x))$ and the actual cases (y_i). Here the $f(x)$ (denotes the number of cases with x, where the x is in the time in number of days from the first case) is given as:

$$f(x) = k.\gamma.\beta.\alpha^{\beta}.x^{-(\beta+1)} exp\left[-\gamma\left(\frac{\alpha}{x}\right)^{\beta}\right], \qquad x, \alpha, \beta, \gamma > 0$$

This can be done using the popular Machine Learning technique of **Levenberg-Marquardt (LM)** for curve fitting as various researches have suggested that during the early periods of COVID, the data had many outliers and noise which makes it hard to accurately predict the number of cases. The base paper has suggested an algorithm that uses an iterative weighting technique called" Robust Fitting" (which is found to be

effective). The main idea is as follows. We maintain weights for all data points (i) in every iteration (n, starting from 0) as w_i^n. First, we fit a curve using the LM technique with weights of all data points as 1, thus $w_i^0 = 1 \; \forall \; i$. Second, we find the weight corresponding to every point for the next iteration) as:

$$w_i^{n+1} = \frac{\exp\left(1 - \frac{d_i^n - \tanh(d_i^n)}{max_i d_i^n - \tanh(d_i^n)}\right)}{\sum_i \exp\left(1 - \frac{d_i^n - \tanh(d_i^n)}{max_i d_i^n - \tanh(d_i^n)}\right)} \quad Iterative\ weighting\ technique: curve\ fitting$$

Simply, in the above equation, we first take **tanhshrink** function defined as **tanhshrink(x) = x − tanh(x)** for the distances of all points along the y-axis from the curve (d_i). This is to have a higher value for points far from the curve and a near 0 value for closer points. This is then standardized by dividing with a max value overall points and subtracted from 1 to get a weight corresponding to each point. This weight is then standardized using the **softmax** function so that the sum of all weights is 1. The curve is fit again using the LM method, now with the new weights (w_i^{n+1}). The algorithm converges when the sum total deviation of all weights becomes lower than a threshold value.

This Levenberg-Marquardt algorithm can be extended to get an approximate fit of various distributions and finding the best-fit parameters corresponding to them. The base paper has proposed that the **Inverse Weibull** function fits the best to the COVID-19 dataset (Overall). We will try fitting iterative versions of Gaussian, Beta (4-parameter), Fisher Tippet, and Lognormal functions. The loss functions which we will be using in order to find the best fit will be MSE (Mean Squared Error), R2 (R-Squared), and MAPE (Mean Absolute Percentage Error).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \qquad MSE = \frac{1}{n}\sum(y_i - \hat{y})^2 \qquad MAPE = \frac{100\%}{n}\sum\left|\frac{y_i - \hat{y}}{y_i}\right|$$

References:

[i] [BASE PAPER] - Predicting the growth and trend of COVID-19 pandemic using Machine Learning and Cloud Computing.

[ii] Background paper on COVID-19 - WHO

[iii] Das S. Prediction of COVID-19 Disease Progression in India: Under the Effect of National Lockdown.

[iv] Predictions, role of interventions and effects of a historic national lockdown in India's response to the COVID-19 pandemic: data science call to arms.

[v] Ghosh, Palash & Ghosh, Rik & Chakra borty, Bibhas. (2020). COVID-19 in India: State-wise Analysis and Prediction

[vi] WHO is investigating reports of recovered COVID patients testing positive again - Reuters.

[vii] The Levenberg-Marquardt Algorithm: Implementation and Theory