Analysis of Covid19 in India

Bhargavi G
Computer Science
PES University
Bangalore, India
bhargu2000@gmail.com

Nisha S
Computer Science
PES University
Bangalore, India
nishasuresh2001@gmail.com

Trisha C Shekhar
Computer Science
PES University
Bangalore, India
trisha.c.shekar5@gmail.com

Nikhil D.B
Computer Science
PES University
Bangalore, India
ndb4263@gmail.com

***Abstract*—Corona virus which originated in a small local seafood market in Wuhan has taken over the whole world. This pandemic has impacted lives in unprecedented ways. In mere nine months Corona virus has spread to over 213 countries and territories. Its spread has left businesses around the world counting costs and wondering what recovery could look like. Many people have lost their jobs or seen their incomes cut due to the coronavirus crisis. Unemployment rates have been on the increase.**

I. Introduction

COVID-19, a novel coronavirus, is currently a major world-wide threat. It has infected more than a million people globally leading to hundred-thousands of deaths.In that reference, it is extremely crucial to construct models that are computationally competent as well as realistic so that they can help policy makers, medical personals and also general public. Modelling the disease and providing future forecast of possible number of daily cases can assist the medical system in getting prepared for the new patients. The statistical prediction models are useful in forecasting as well as controlling the global epidemic threat.Even as this summary is written the number of new cases are increasing around the world. We thought covid19 analysis would be the best topic to work on as we ourselves can get to know its impact on Indian economy and its people. We decided to restrict ourselves to analyse the spread of coronavirus in India only.

A number of papers related to the analysis of coronavirus were read and the following papers were considered.

A. Tracking the Spread of COVID-19 Cases in India using Data Visualizing and Forecasting Techniques

Visualization:

Python libraries like matplotlib and plotly are used for data visualisation. Bar charts, gantt charts among others are used to represent different features of the covid19 dataset. A number of visualisations have been shown in the paper. Some of the visualisations mentioned are, a bar chart to represent the number of active cases in each state, a horizontal bar graph to show the position of India in the world in terms of the number of cases recorded and a chart to represent the number of active, recovered and deceased people in each state. Analysis is not restricted to Indian states.

Identify applicable funding agency here. If none, delete this.

Visualisations in this paper are based on data collected till the end of April but we are using data collected till the end of September.

Figure 1 is one of the data visualizations from the paper. It is a bar chart showing the top 15 countries along with the number of covid cases reported in those countries.

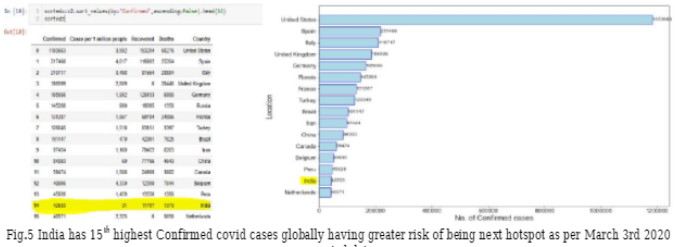
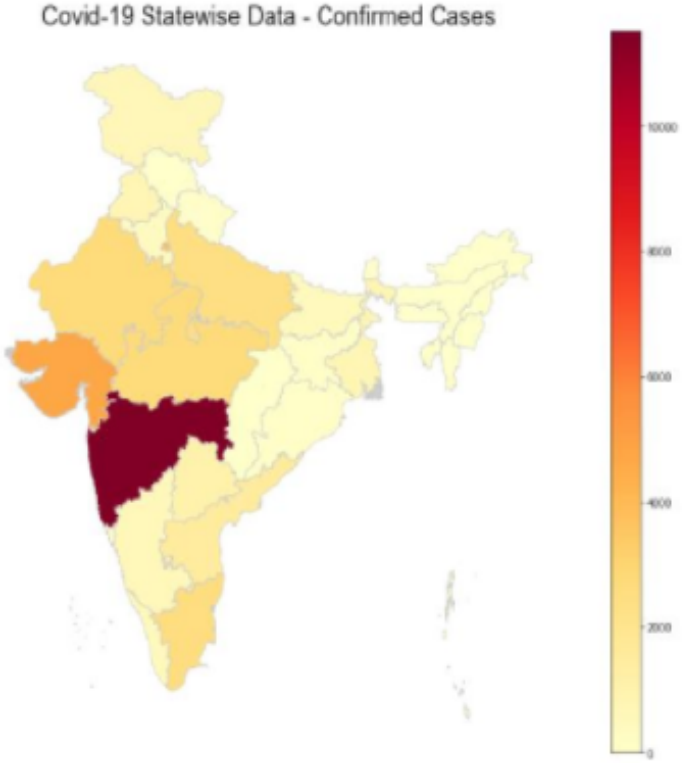


Fig.5 India has 15th highest Confirmed covid cases globally having greater risk of being next hotspot as per March 3rd 2020 reported data

Fig. 2 shows the visualization of choropleth maps for Indian states with respect to confirmed covid cases.



Forecasting:

The ARIMA model is used for forecasting. In the ARIMA

model Dickey-fuller test is used to test the stationarity of time series. Autocorrelation function and partial autocorrelation functions are used to determine parameters for this ARIMA model. The parameters found are then passed to the ARIMA model for forecasting the actual time series for confirmed covid cases across dates in India. According to the values predicted using the ARIMA model it is observed that the number of cases increases exponentially over time.

Holt winter model is also used for forecasting. It is simpler in terms of implementation when compared to ARIMA and has almost the same accuracy as ARIMA. This model also predicted that the number of cases would increase over time.

B. Time Series Prediction Based on Linear Regression and SVR

This paper uses linear regression and support vector regression for time series analysis. They are combined and not used separately because the combined model gives better efficiency than when individual models are used. Time series is divided into linear and non-linear parts. This linear part is the stable part of the time series and the non-linear part is the unstable part of time series which contains some irregular variation. First-order linear model as a stable part. The two parts, linear and non-linear of this new model are predicted separately and then integrated together for forecasting. The new regression implemented was better than common SVR 40% in precision of forecasting.

Pre-processing of time series is done by splitting the time series by first order linear regression. Pre-processing of time series helps in forecasting the two parts mentioned earlier. To do first order linear regression for a time series x t where t=1,2,...,N the following formulas were used.

$$b = \frac{\sum x_i t - n\bar{x}\bar{t}}{\sum t^2 - n\bar{t}^2}$$

$$a = \bar{x} - b\bar{t}$$

Where, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t$

Once a and b are found the linear part of the time series x' is set to a+bt and the nonlinear part x" is set to x-x'. Pre-processing of time series is followed by phase space reconstruction and linear and SVR forecasting. The model is finalised by selecting those values for the parameters which give better accuracy. The following table is the result of this paper which shows that the precision result of the new model is better than just SVR.

Table 1 prediction data between two manners

Source data	41.5652	34.179	48.0958
SVR data	45.0366	36.2981	36.8933
New model data	45.4356	34.1775	35.7289
Source data	46.3303	38.39649	39.6433
SVR data	37.1143	35.831	36.5036
New model data	44.5543	41.7416	37.7681

C. Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future

Corona virus (COVID -19) is an ongoing pandemic that has infected more than 39.7 million people, taking away life's of more than 1.11 million people. In that reference it's important to construct competent and realistic models which help us to analyse present situation, forecast future number of cases, assist to take precautionary measures and control the pandemic. In this paper they have employed Auto-Regressive Integrated Moving Average (ARIMA) model for prediction.

Data: Confirmed, recovered and death cases of COVID-19 infection are collected for India as well as countries with highest confirmed and countries in South-East Asia region, as per World Health Organization region classification, from the official website of Johns Hopkins University (https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html) from 22 January 2020 to 13 April 2020.

Assumptions: In this paper they have followed normality assumption with residual histogram backs up the assumption. The model satisfies assumption of constant variance. We can also observe that the p-value of all the lags is larger than the significance level (i.e. 0.05) which means there is no violation of independent assumption.

Tech: ARIMA models and parameters is expressed as ARIMA (p, d, q) where p stands for the order of auto-regression, d signifies the degree of trend difference while q is the order of moving average. To find the initial number they use Autocorrelation function (ACF) graph and partial autocorrelation (PACF) graph. Then they test the models for variance in normality and stationary. To determine the best suited model for forecast is done by observing their MAPE, MAD and MSD values. The finest model has the lowest value for all the measures. The best fit ARIMA model is compared with Linear Trend, Quadratic Trend, SCurve Trend, Moving Average, Single Exponential, Double Exponential models. Then the built model is employed to forecast confirmed COVID-19 cases for the next 20 days. The model for forecasting future confirmed COVID-19 cases is represented as, ARIMA(p,d,f):

$$Xt = \alpha 1 Xt-1 + \alpha 2 Xt-2 + \beta 1 Zt-1 + \beta 2 Zt-2 + Zt$$

where $Zt = Xt - Xt-1$

Here, Xt is the predicted number of confirmed COVID-19 cases at t th day, alpha1, alpha2, beta1 and beta2 are parameters whereas Zt is the residual term for t th day.

To verify the efficiency of model and to get an idea of recovery and death trends they have set the level of statistical significance at 0.05 and a graph is plotted for actual confirmed cases and predicted confirmed cases with respect to time. They have also performed a comparative study to examine the status of confirmed COVID-19 cases of India with respect to those of highly infected countries and south -east Asian region

Main claims: The results for measure of model accuracy suggests that ARIMA (2,2,2) is most accurate for forecast because of its least value for all the measures. The AR (2) and MA (2) estimates p-values imply that parameters are significant in the model. We can observe a slight deviation of residuals from the straight line which indicates that there aren't many errors but we have a few outliers in the dataset. They also claim that there is high rise in cases plotting the time series graph. They check for precision of model in forecasting by plotting actual confirmed cases vs forecasted cases. They claim that the disease has low mortality because the rate of recovery is high. They suggest that is India is one among the most infected countries.

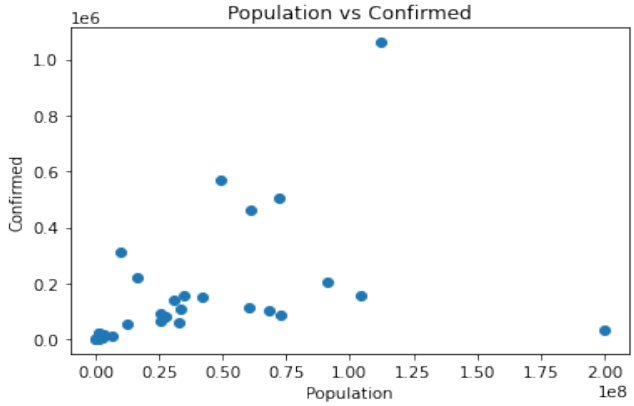
Conclusion: A look at the MAPE, MAD and MSD values suggests that ARIMA (2, 2, 2) model is the most accurate of all for forecasting. As compared to other prediction models, for instance support vector machine (SVM) and wavelet neural network (WNN), ARIMA model is more capable in the prediction of natural adversities. Although a large amount of data helps in providing a more exhaustive prediction and explanation, in the present circumstance, these models could be valuable in anticipating future cases of infection if the pattern of virus spread didn't change abnormally. It is obvious that this virus is new and has the capability to be transmitted intensely. Hence, it may have an influence on the predictions. The time series analysis shows an exponential enhancement in the infected cases. A comparative study with some of the highly infected countries and countries in south-east Asia region indicates that India was the last amongst these countries to get infected.

II. PROPOSED PROBLEM STATEMENT

With this project, we aim at analysing the current covid19 situation in India and predicting the trend in the rising cases.We also take into many other factors that might have relations with the increase in cases in each state. We take into account 7 data sets, perform Exploratory Data Analysis on various problems and answer the key problem statements related to them.

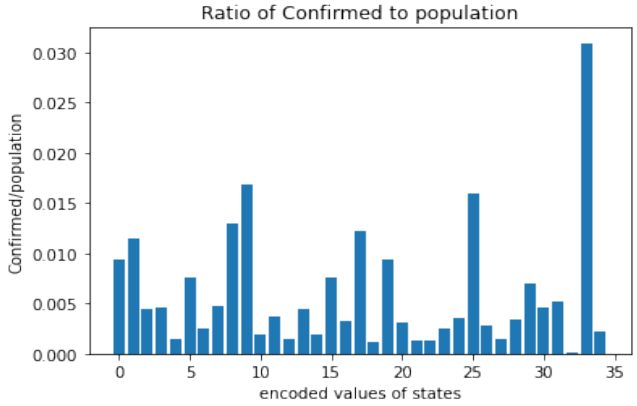
A. Extract population and state/ union territory, Analyse the percent of active or total covid cases with respect to the population of each state and to check if the data follows a particular trend.

Logically one might infer that states with larger population have higher number of confirmed cases compared to the others, but on analysis of the relationship between total Population and covid cases, we got a Pearson's correlation of 0.445.



From the above graph, we see a positive correlation

We have also analysed how each state is performing and figuring out which states have contained the spread well compared to the others. Here, we take in the ratio of the confirmed cases to the population for the comparison.

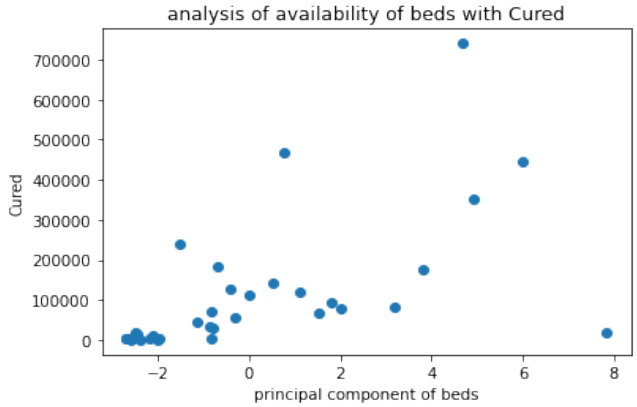


Here, we find that ratios are high with places of smaller population. For example Uttarakhand, Goa and Puducherry. We do know that higher the population, more difficult it is to contain the spread of virus in the state, therefore states with lesser ratio are doing a better job despite their huge population

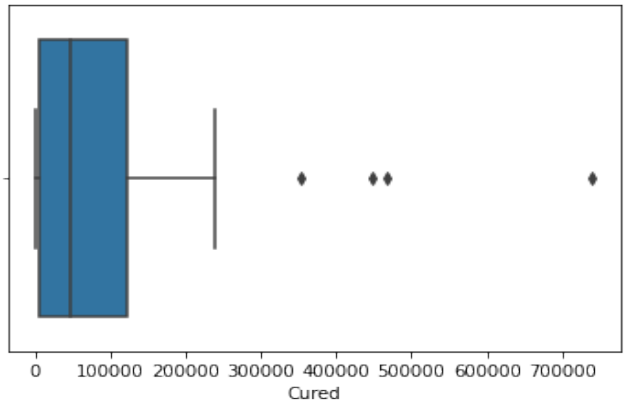
B. Analyse the availability of hospital beds with the recovery rate

Our Beds data set contains the number of urban beds,rural beds and public beds available along with the number of community,private,district and sub-district health cares available for each state. This dataset contains 6 NaN values in one column. We have filled the NaN values with the mean of the column as the column contains continuous values. We are assuming that all the columns are uncorrelated and analysing all these columns with respect the cured rate for

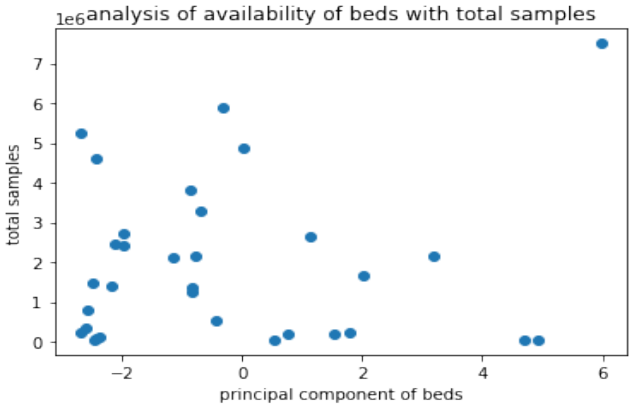
each state.As the features for analysis is multi-dimensional, we must perform dimensionality reduction for further analysis. We use Principal Component Analysis to acheive the same. We find that the correlation between Principal Component value and the recovered values is 0.588.



We take the recovered cases from the covid 19 dataset where we find the the cured column is positively skewed containing outliers.

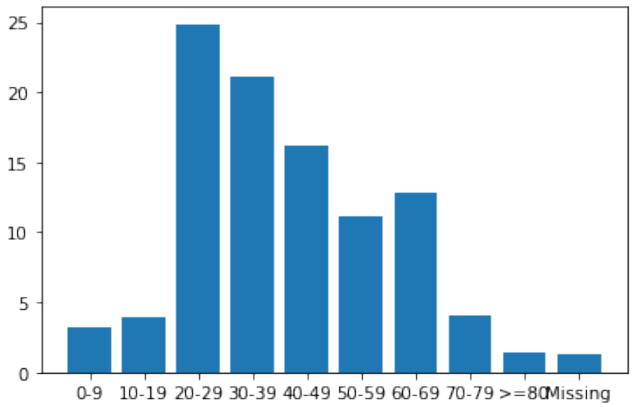


We also included the visualizations for the relation between the principal component and the total samples. Here, we find that two columns are not correlated at all with a Pearson correlation coefficient of 0.096. This explains the strong relationship between the increase in cured cases with the increase in facilities in all the states



C. Analyse age groups at higher risk and percentage affected

The Age dataset contains the age group, total cases and percentage. We have 9 missing values which have been taken care of. Here we find that the age group of 20-29 are affected the most. 25% of the total covid cases are from this group.

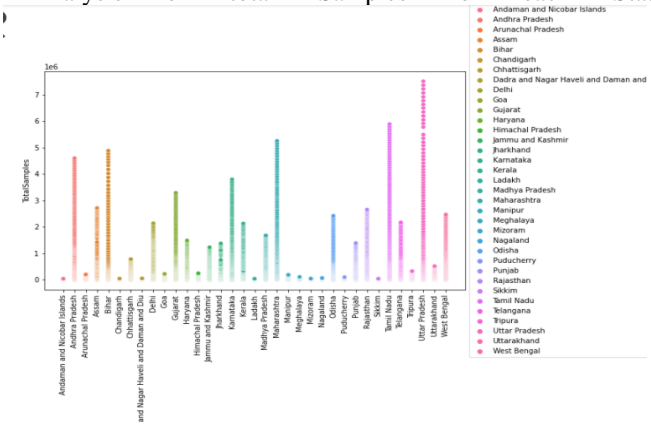


We also have a dataset that analyses each and every individual that has contracted the virus in the beginning stages. Not much analysis can be done as most of the details of the patient history are not available. But from this data, we do gather that most of the individual cases are from the state of Maharashtra in the beginning stages. The same result was expected.

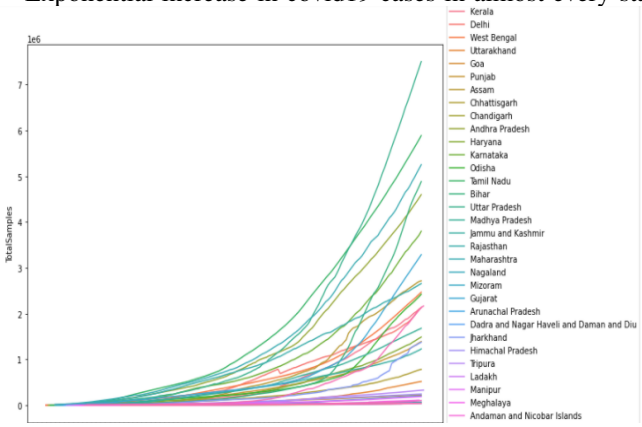
D. Analysis of Covid 19 cases per State

While analysing the State Dataframe, we found that the number of positive cases and the number of negative cases do not add up to the total samples given in the dataframe. This made us doubt the correctness of the dataset. We analysed the trend in the total samples of each state and found that the graph is exponential growth in the total samples from April 1st 2020 to September 14 2020. Some areas like Megalaya,Sikkim,Puducherry,Himachal Pradesh, Arunachal Pradesh etc. remained constantly low in covid samples till the very end but places like Odhissa, Karnataka, Andhra Pradesh show an exponential growth in cases. We find Uttar Pradesh having the highest number of samples and the islands having the lowest.

Analysis of total Samples for each State:

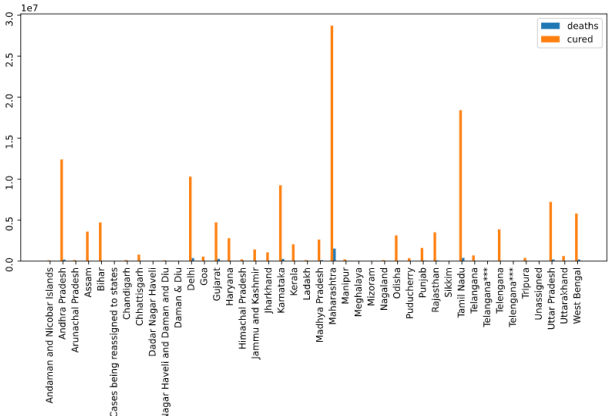


Exponential increase in covid19 cases in almost every state



We plan on using the dataset on which the above graph was plotted for predicting the trend in the rising cases in each State. We plan on estimating through many approaches including ARIMA,SVMs and Regression by re training on our dataset at specific intervals of time. We have considered various other factors that are small but effective in providing some insight to the rising cases in India. Our approach for the predictions will be based on the model that gives the best outputs.

Analysis of the Death vs Cured for each state



III. References:

[1]https://www.researchgate.net/publication/341297682$_Tracking_the_Spread_of_COVID-19_Cases_in_India_using_Data_Visualizing_and_Forecasting_Techniques$

[2]https://arxiv.org/ftp/arxiv/papers/2004/2004.07859.pdf

[3]https://core.ac.uk/download/pdf/41373796.pdf

[4]https://www.bbc.com/news/business-51706225