Data Mining Project Report

COVID-19 Analysis and Prediction

Group No. : 27
Rohit Raj (20111051) rohitjha20@iitk.ac.in
Shruti Wasnik (20111062) gdshruti20@iitk.ac.in
Sonam Tshering (20111064) somsring20@iitk.ac.in
Mani Kant Kumar (20111030) mani20@iitk.ac.in

Project Instructor:
Associate Prof. Arnab Bhattacharya

Master of Technology
Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur
December, 2020

Contents

1 Acknowledgement

We would like to express our sincere gratitude to project instructor Associate Prof. Arnab Bhattacharya for providing their indispensable guidance, comments, suggestions throughout the course of the project. He has been truly supporting and helped us whenever we were in need. We thank him for allotting some precious time from his busy schedule without which it would be very hard for us to proceed in the project.

2 Abstract

We have been severely affected by the pandemic from a long time with so many efforts being made to control the spread and discover the vaccine, we still cannot predict when the things will get back to normal. We never know what will happen if we are not careful. The COVID-19 cases in India have increased abruptly few months back and the research of vaccine is still under testing phases. It may take a long time to supply a safe and effective vaccine to common people at affordable price. It is well known fact that *"Prevention is better than cure"* which implies it's easier to stop something happening in the first place than to repair the damage after it has happened.

It has become more important for us to be more careful, and follow the preventive measures to avoid the disease. So we the team of 4 enthusiastic students worked on a project in which we had examined the data of COVID cases to understand the pattern of spread in different regions of India and used those analysis to predict cases for the next few days according to the recent spread pattern which will certainly give an insight to the seriousness of the pandemic.

3 Problem Statement

We aim to analyze the number of COVID-19 infected people in India starting from the day when first case was detected till now and present these analyse data in superficial way. The user can fetch data of specific district, state or can fetch the whole data for a specific date or month. This analysis includes

- number of confirmed cases
- number of recovered cases
- number of deceased cases
- age bracket more prone to COVID-19
- gender-wise mortality pattern(death vs hospitalized)
- prediction of new cases

To give these statistics a better perception, we intent to plot these data. Moreover, these analysed data is utilized to predict the number of cases in the near future.

4 Introduction and Motivation

The highly infectious coronavirus disease (COVID-19) was first detected in Wuhan, China in December 2019 and subsequently become epidemic, infecting millions of people. In India, a large country of about 1.3 billion people, the disease was first detected on January 30, 2020, in a student returning from Wuhan.

Most of the prior research focused on the number of infections in the entire country. However, given the size and diversity of India, it is important to look at the spread of the disease in each state and district separately, wherein the situations are quite different. So we intent to build a model which can provide the analysis of the pandemic outbreak situation in India, the rising trend of COVID-19, hot-spots and cold-spots, and the mortality rate. This information can be further used for making future predictions. Since the cases are rising very fast, some aggressive and effective control strategies are required to prevent the transmission. This analysis will help government of different states to plan suitable strategies to tackle the situation. It will create awareness among people to stay safe and take preventive measures. Some of the benefits of this analysis are:

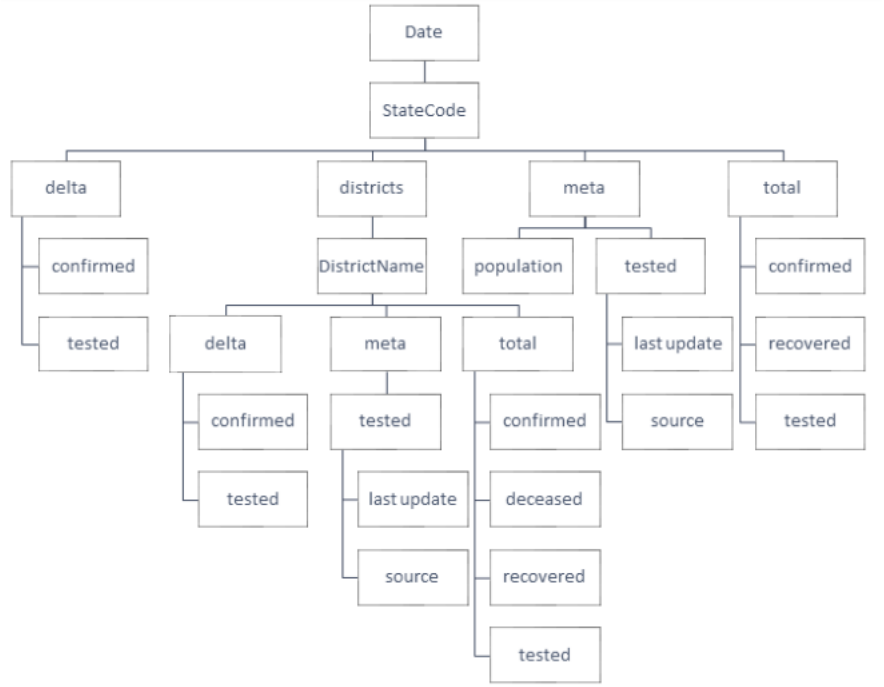
- The analysis could be used for finding the hot-spots, so that more attention can be provided to those regions.
- The mortality rate by age data can be used to analyze the age group which are more prone to the spread. Special medical facilities and preventive measures can be provided to them.
- The predicted cases can be used by the hospitals to arrange extra vacant beds and medical equipment in advance, such that treatment can be provided to everyone.

5 Datasets used

Since the data of COVID-19 cases is updated periodically, so we are fetching this dynamic data online from the website `https://www.covid19india.org/`. The APIs needed to access the data are available at `https://api.covid19india.org/`. The source files are available in different forms like district-wise, daily changes, time series, state-wise stats, and raw data. The information of patient wise cases is available in the raw-data files which are mined for calculating the mortality rate by age. Out of many forms and formats, we mainly use National/State/District Level: Historical date-wise cumulative/daily counts from the data-all.json file. Since, it includes cases for each date, the analysis for respective time period can be done more conveniently. We extracted the necessary data from these files.

- `https://api.covid19india.org/csv/latest/district_wise.csv` contains district-wise data.
- `https://api.covid19india.org/v4/data-all.json` includes day-wise updates on cases.
- `https://api.covid19india.org/raw_data19.json` contains patient-wise data.
- `https://un-mapped.carto.com/tables/states_india/public/map` supports mapping of cases on map of India.

The structure of the data available in data-all.json is represented by the following diagram:



6 Methodology

Python programming language is used for building this project, since it provides a great ecosystem for data science. It includes a number of supportive libraries which can be used for data pre-processing, modelling, manipulation, visualization, prediction and for other processes as well.

6.1 Pre-processing

We have used data pre-processing and data selection techniques to prepare the data and `pandas` to store the dataframe extracted from the source files. This pre-processing includes data cleaning, handling missing data, data integration, etc. For instance suppose we have three fields: "cases", "deceased" and "recovered". If any of the three fields data is available on a specific day then we took this data into consideration by assigning the remaining field data as zero.

6.2 Visualization

Data Visualization method is used to present the prediction and analysis work by means of graphs, charts, and `choropleth maps, seaborn` which made it easier to detect patterns, trends, and outliers in the data. We imported `ipywidgets, IPython.display` etc. libraries to take input from the people without having much technical skill. HTML is used to represent the analysis in simplistic and attractive way. We used `matplotlib.pyplot` to plot the analysed data so that user can have better visualization.

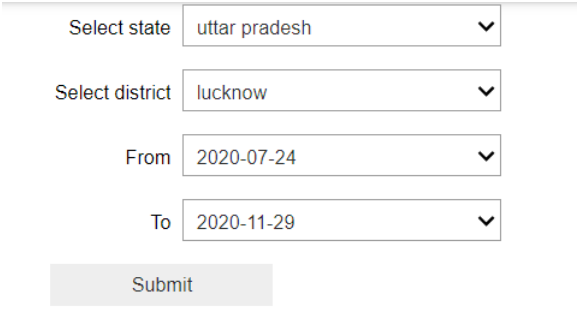
6.3 Prediction

We used regression models to analyze the trend of spread and built a model which is used for predicting the cases based on the recent trends.We tested our ideology on various regression models, like *Linear Regression, Multiple Regression, Polynomial Regression, Exponential Regression, etc.*, but the linear model seems to give the best accuracy.

7 Results

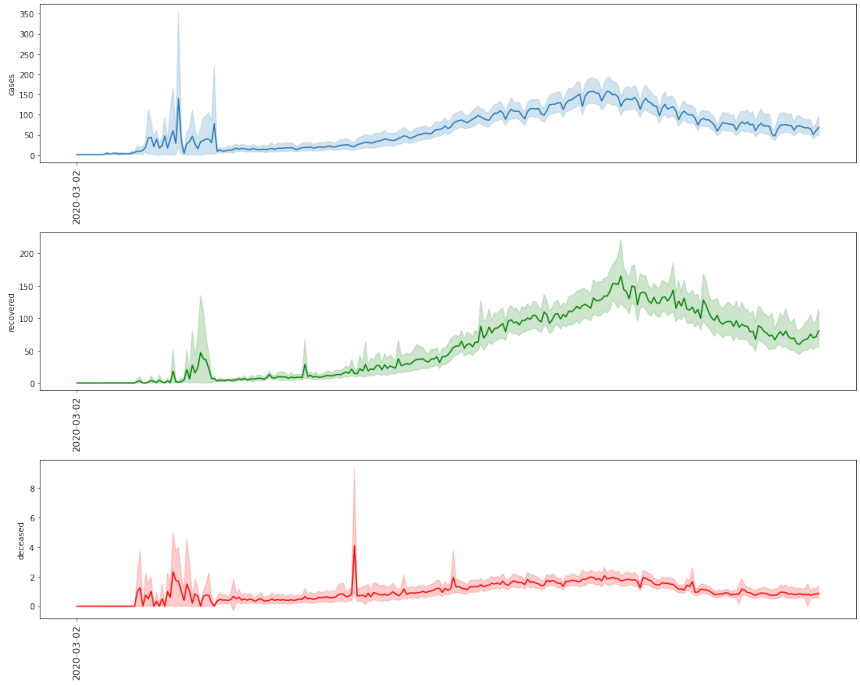
The analysis of COVID-19 spread can be visualized for all districts and states of India in a particular time-series. The model computed the number of total confirmed cases, total deceased cases and total recovered cases from beginning till now.

We tried to provide GUI like interface to take input from users. One can choose states, districts and dates from the dropdown menu as shown below:

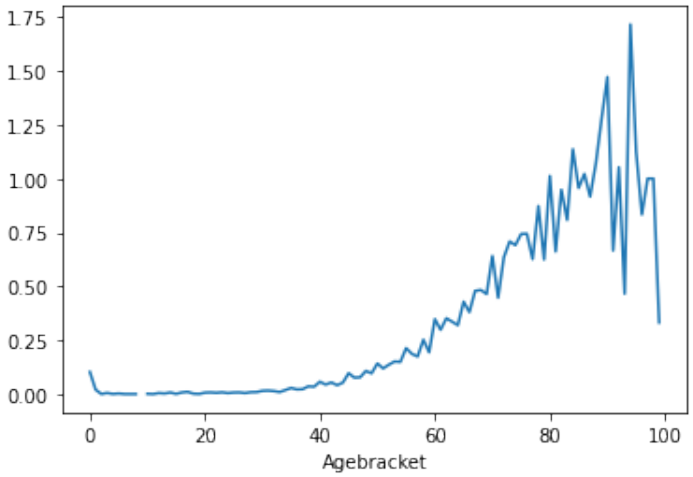


7.1 Analysis

We can visualize the elevation and depression in the confirmed cases, recovered cases, and deceased cases as plotted below:

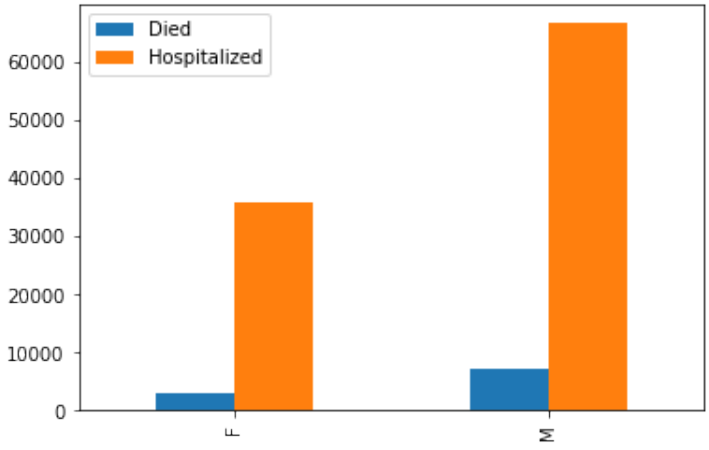


The model depicts the age bracket which is more prone to death. The plot for the same can be visualized as:



The model also depicts a bar graph of died vs hospitalised patients as plotted below.

- **F**: represents female patients
- **M**: represents male patients



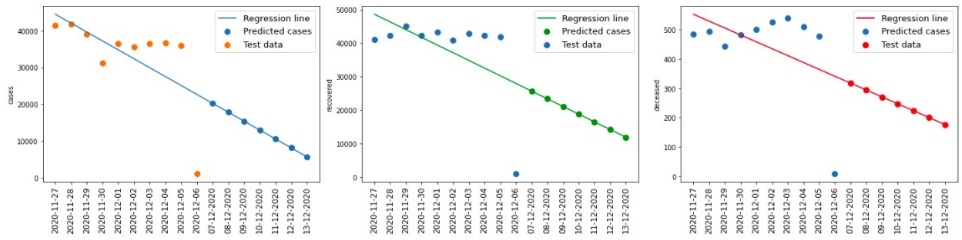
7.2 Prediction:

The data set has been divided into two parts -the training set and the validation set. The model is trained based on previous consecutive 10 days data as training data, while the test data will be next 7 consecutive days for which we want prediction. The predicted model output and graph for next week is shown in figure below:

date	cases	deceased	recovered
06-12-2020	19518	325	26489
07-12-2020	16833	303	24376
08-12-2020	14148	280	22264
09-12-2020	11463	258	20152
10-12-2020	8777	236	18039
11-12-2020	6092	213	15927
12-12-2020	3407	191	13815

Predictions for state : ALL and district : ALL

Prediction graphs

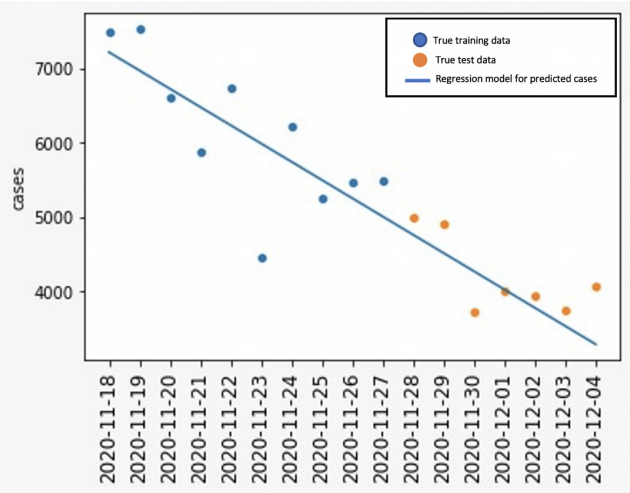


7.3 Prediction Evaluation:

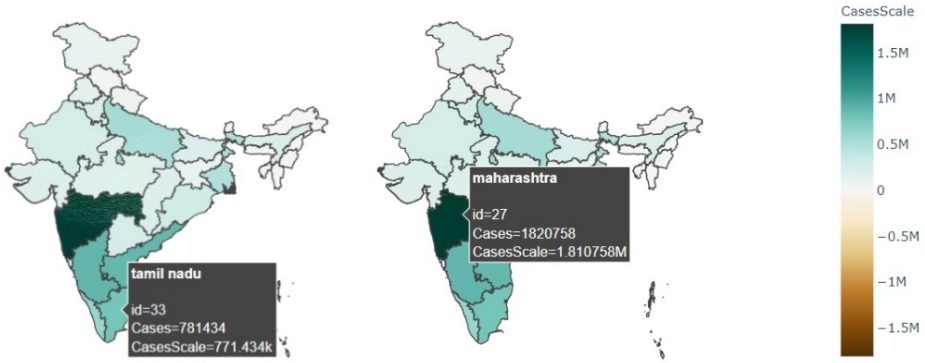
The accuracy of the model is evaluated based on model output that we predicted and the actual data that we already have from web resources. The predictions are made based on the number of cases found just a few days back. We tested on various models, and linear regression seems to give the best accuracy. The predicted data is closed to the actual data as shown in table below:

	date	predicted cases	actual cases
0	2020-11-28	4761	4998
1	2020-11-29	4515	4906
2	2020-11-30	4269	3726
3	2020-12-01	4023	4006
4	2020-12-02	3777	3944
5	2020-12-03	3530	3734
6	2020-12-04	3284	4067

We can also visualize it from the graph. We can notice that yellow points lie close to blue line which implies predicted output is very much close to real output.



8 Discussion



The intensity of colour shows the hotspot and coldspot regions in India. We can conclude from the above map that Maharashtra is the most affected among all states in India. It is found that five states namely, Maharashtra, Karnataka, Andhra Pradesh, Tamil Nadu and Kerala are hotspots whereas Dadra and Nagar Haveli and Daman and Diu, Mizoram, Andaman and Nicobar, Sikkim, Ladakh are coldspots. The COVID-19 virus infects people of all ages. However, evidence to date suggests elderly people (mostly above 60) are at a higher risk of getting severe COVID-19 disease due to their decreased immunity and body reserves.

Hotspots

state	cases	deceased	recovered
maharashtra	1834519	47257	1708862
karnataka	889857	11815	853279
andhra pradesh	869578	6989	857002
tamil nadu	785669	11738	763834
kerala	625299	2355	561532

Coldspots

state	cases	deceased	recovered
dadra and nagar haveli and daman and diu	3334	2	3289
mizoram	3912	6	3690
andaman and nicobar islands	4742	61	4611
sikkim	5147	111	4587
ladakh	8712	120	7734

9 Future Direction

In this project we are analysing the data only for India. However, the same model can be build for other nations as well. For now we are able to make prediction for next few days but we will try to predict for more weeks. We can enhance the accuracy with flexible algorithms. We can predict new cases by dynamically choosing the algorithms like linear regression or polynomial regression etc. whichever suits the most, depending upon the previous norms. We will make a web interface to interact with users in convenient way while taking inputs and will show all the analysed data in superficial way.

10 Instructions to run

The code can be run on jupyter notebook or google collab. We need to import a file named `states_india.geojson` on the platform which is available in zip file.

10.1 Libraries imported

- from urllib.request import urlopen
- import ipywidgets as widgets
- from *IPython.display* import display, clear_output,Markdown,HTML
- import *matplotlib.pyplot* as plt
- import seaborn as sns
- import plotly.express as px
- import plotly.io as pio
- import seaborn as sns
- from sklearn.model_selection import train_test_split
- from sklearn.linear_model import LinearRegression

10.2 Installations

- Libraries: seaborn, matplotlib, plotly, sklearn, datetime.