COVID-19 Pandemic & Machine Learning

Khushbu Mevada (MIT2019086) Shivam Kasat (MIT2019024)

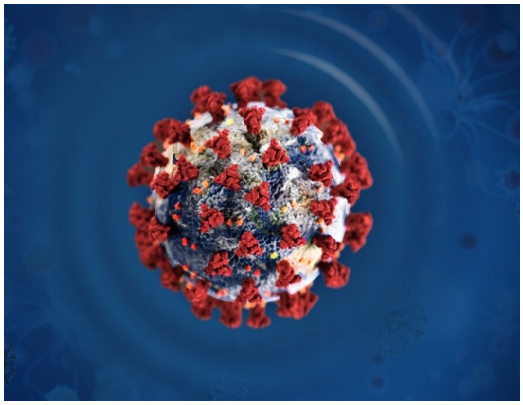
04 May 2020

Abstract

The current research proposes a model to predict the number of people that may be affected, the peak period and the halt time for the coronavirus pandemic outbreak worldwide. Here, we apply Machine learning models to predict aforementioned predictions. We use the statistical data generated from countries, namely India, Spain and South Korea and we also use the statistical data from some of the Indian states for prediction purposes in order to get some meaningful insights from our analysis.

Introduction

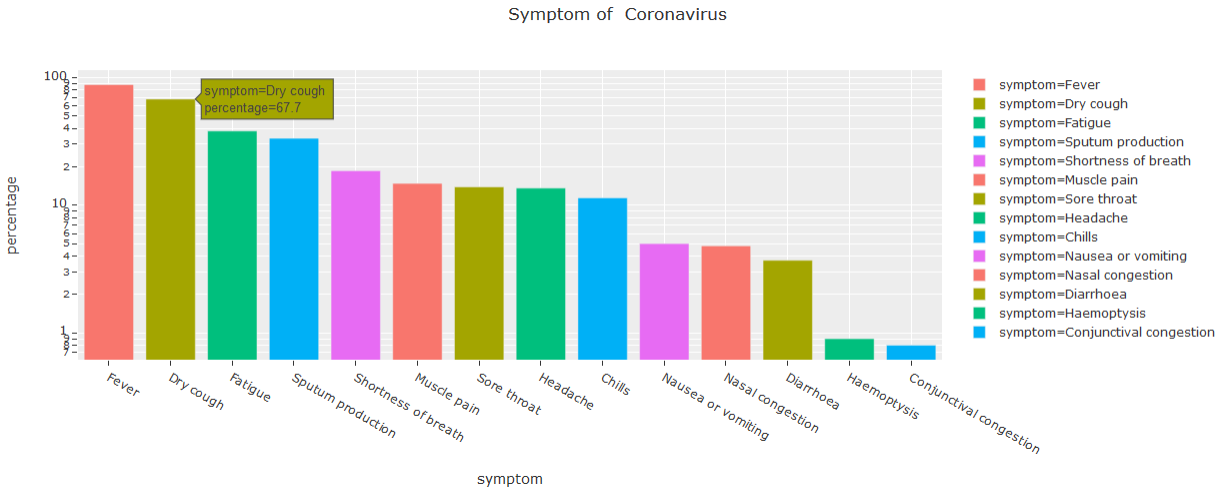
What is coronavirus?
Coronaviruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19.



COVID-19 is the infectious disease caused by the most recently discovered coronavirus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019.

The first case of 2019-2020 coronavirus pandemic in India was reported on 30^{th} January, 2020, Originating from China. Experts suggest the number of infections could be much higher as India's testing rates are among the lowest in the world. The infection rate of COVID-19 in India is reported to be 1.7, significantly lower than in the worst affected countries.

Symptoms of Coronavirus



The above graph represents the probability of having corona positive given the symptom. e.g., There is 67.7% probability for a person to be tested positive given he has dry cough as a symptom.

Datasets

We have used multiple dataset for our analysis and prediction task, which were taken from kaggle.com

- **COVID19 GLOBAL Forecast**

  The White House Office of Science and Technology Policy (OSTP) pulled together a coalition research groups and companies (including Kaggle) to prepare the COVID-19 Open Research Dataset (CORD-19) to attempt to address key open scientific questions on COVID-19. Those questions are drawn from National Academies of Sciences, Engineering, and Medicine's (NASEM) and the World Health Organization (WHO).
  It has six columns named Id, Province state, Country Region, Date, Confirmed Cases, Fatalities.
  They are currently updating the data daily.

- **COVID-19 in India**

  This dataset has information from the states and union territories of India at daily level. State level data comes from Ministry of Health & Family Welfare. Individual level data comes from **covid19india.**

  - COVID-19 cases at daily level is present in covid_19_india.csv file.
  - Individual level details are present in IndividualDetails.csv file.
  - Population at state level is present in population_india_census2011.csv file.
  - Number of COVID-19 tests at daily level in ICMRTestingDetails.csv file.
  - Number of hospital beds in each state in present in HospitalBedsIndia.csv file.

- **Novel Corona Virus 2019 Dataset**

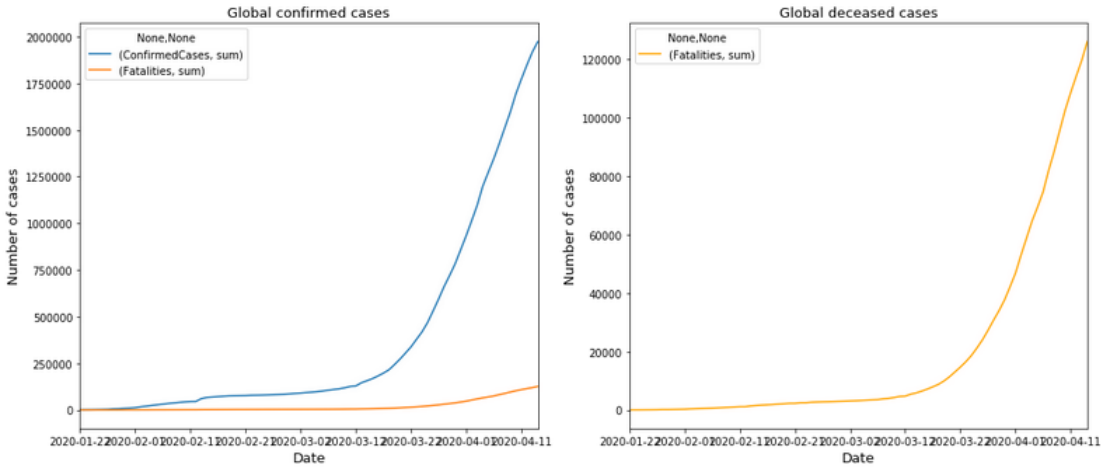
This dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. Please note that this is a time series data and so the number of cases on any given day is the cumulative number.
Main file in this dataset is covid_19_data.csv and the detailed descriptions are below.

- Sno - Serial number
- ObservationDate - Date of the observation in MM/DD/YYYY
- Province/State - Province or state of the observation (Could be empty when missing)
- Country/Region - Country of observation
- Last Update - Time in UTC at which the row is updated for the given province or country.
- Confirmed - Cumulative number of confirmed cases till that date
- Deaths - Cumulative number of of deaths till that date
- Recovered - Cumulative number of recovered cases till that date

Exploratory Data Analysis (EDA)

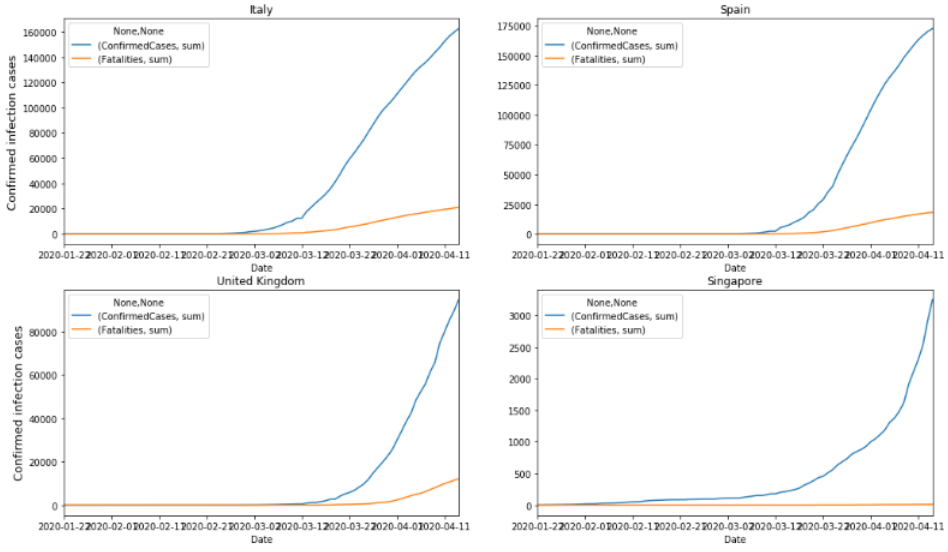
EDA on COVID19 GLOBAL Forecast dataset

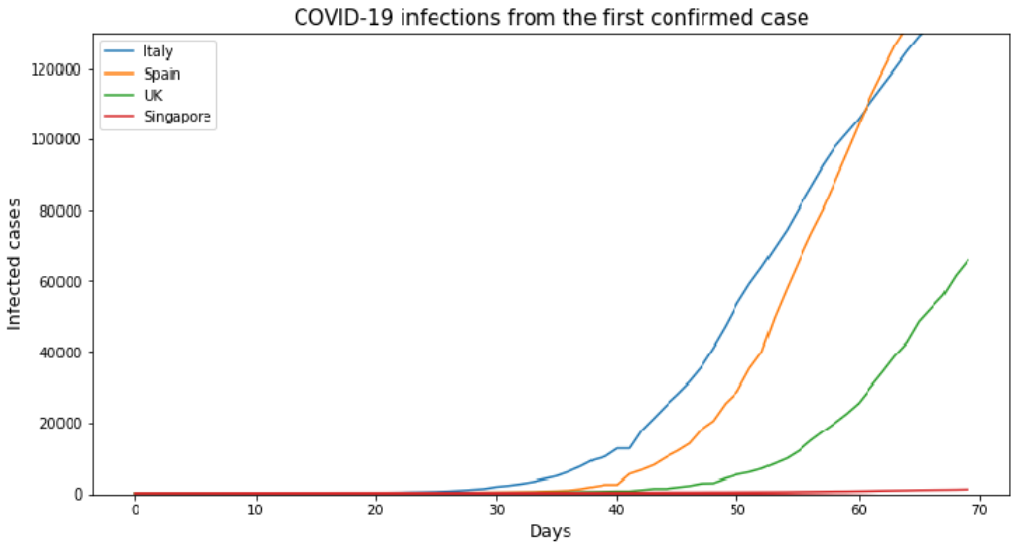
- The dataset covers 163 countries and almost 2 full months from 2020, which is enough data to get some clues about the pandemic. We will see a few plots of the worldwide tendency to see if we can extract some insights:



Observation:
We can see that the global curve shows rich fine structure, but these numbers are strongly affected by the vector zero country, China. Given that COVID-19 started there, during the initial expansion of the virus there was no reliable information about the real infected cases.

- Both Italy and Spain are experiencing the larger increase in COVID-19 positives in Europe. At the same time, UK is a unique case given that it's one of the most important countries in Europe but recently has left the European Union. The fourth country we have taken is Singapore, since it is closer to China and its socio-economic conditions is different from the other three countries.





Observations:

- **Italy**
  With almost 120,000 confirmed cases, Italy shows one of the most alarming scenarios of COVID-19 with more than 2% of population has been infected.
- **Spain**
  Spain has the same number of cumulative infected cases as Italy, near 120,000. However, Spain's total population is lower (around 42 millions) and hence the percentage of population that has been infected rises up to 3%.
- **United Kingdom**
  Despite not being very far from them, the UK shows less cases. The number of cases is around 40.000, this is, a 0.6% of the total population.
- **Singapore**
  Singapore is relatively isolated given that is an island, and the number of international

travels is lower than for the other 3 countries. The number of cases is still very low (¡1000) with 0.2% of population being infected.

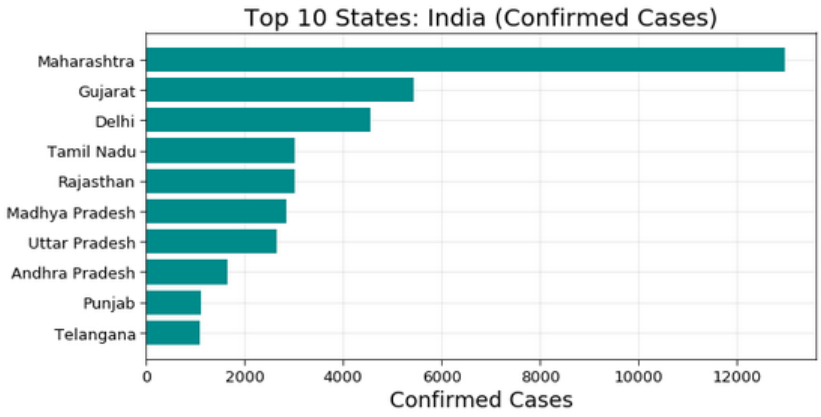
EDA on COVID-19 INDIA dataset

- The following table shows the details like total number of confirmed cases, number of recovered cases, deaths, active cases, mortality rate for all the states/union territories having active cases till date 4^{th} May, 2020.

Out[10]:

state	confirmed	recovered	deaths	active	Mortality Rate (per 100)
Maharashtra	12974	2115	548	10311	4.220000
Gujarat	5428	1042	290	4096	5.340000
Delhi	4549	1362	64	3123	1.410000
Madhya Pradesh	2837	798	156	1883	5.500000
Rajasthan	3016	1356	75	1585	2.490000
Tamil Nadu	3023	1379	30	1614	0.990000
Uttar Pradesh	2645	754	43	1848	1.630000
Andhra Pradesh	1650	524	33	1093	2.000000
Telangana	1082	545	29	508	2.680000
West Bengal	963	151	50	762	5.190000
Jammu and Kashmir	701	287	8	406	1.140000
Karnataka	642	304	26	311	4.050000
Kerala	500	401	4	95	0.800000
Bihar	523	124	4	395	0.760000
Punjab	1102	117	21	964	1.910000
Haryana	463	251	5	207	1.080000
Odisha	163	60	1	102	0.610000
Jharkhand	115	27	3	85	2.610000
Chandigarh	97	19	1	77	1.030000
Uttarakhand	60	39	1	20	1.670000
Himachal Pradesh	40	34	2	1	5.000000
Assam	43	33	1	9	2.330000
Meghalaya	12	10	1	1	8.330000

- Top 10 most affected states of India with confirmed number of cases till date 4^{th} May, 2020.



Prediction

Preprocessing data

- **Join data.**
  Join train/test to facilitate data transformations
- **Filter dates.**
  remove ConfirmedCases and Fatalities post 2020-03-12. Create additional date columns
- **Missing values.**
  Analyze and fix missing values

Compute lags and trends

- **Lag**
  Lags are a way to compute the previous value of a column, so that the lag 1 for ConfirmedCases would inform the this column from the previous day. The lag 3 of a feature X is :

$$X_{lag3}(t) = X(t-3)$$

- **Trend**
  Transformig a column into its trend gives the natural tendency of this column, which is different from the raw value. The definition of trend we applied is:

$$Trend_X = \frac{X(t) - X(t-1)}{X(t-1)}$$

The backlog of lags we applied is 14 days, while for trends is 7 days.

Add country details

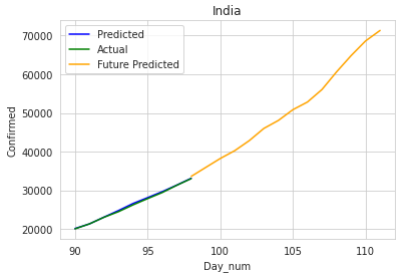
Variables like the total population of a country, the average age of citizens or the fraction of people living in cities strongly impact on the COVID-19 transmission behavior. so, we added those details to data.

Linear regression model for prediction

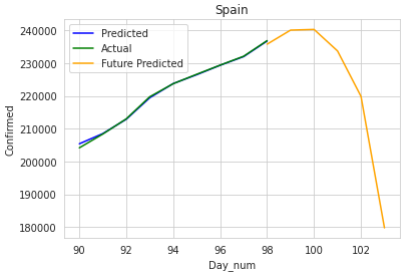
1. **Features.** Select features
2. **Dates.** Filter train data from 2020-03-01 to 2020-03-20
3. Begin with the train dataset, with all cases and lags reported
4. Forecast only the following day, through the Linear Regression
5. Set the new prediction as a confirmed case
6. Recompute lags
7. Repeat from step 4 to step 6 for all remaining days

Observations:

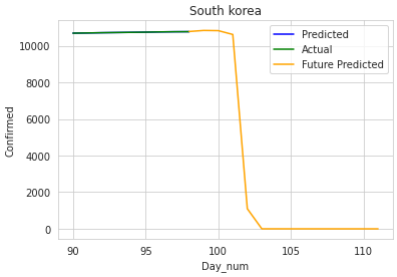
- **India**



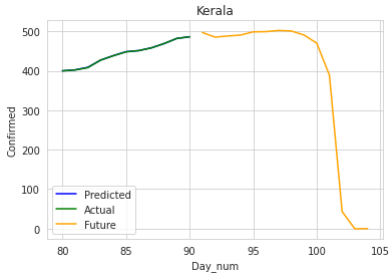
- **Spain**



- **South Korea**



- **Kerala**



Prevention

To avoid the critical situation people are suggested to do following things

- Avoid contact with people who are sick.
- Avoid touching your eyes, nose, and mouth.
- Stay home when you are sick.
- Cover your cough or sneeze with a tissue, then throw the tissue in the trash.
- Clean and disinfect frequently touched objects and surfaces using a regular household
- Wash your hands often with soap and water, especially after going to the bathroom; before eating; and after blowing your nose, coughing, or sneezing. If soap and water are not readily available, use an alcohol-based hand sanitizer.

Conclusion

From the above mentioned graphs we can conclude that according to our linear regression model

- the peak on pandemic in **India** may come after almost 110 days.
- the peak on pandemic in **Spain** may come between 95-100 days and will finished after 105 days.
- the peak on pandemic in **South Korea** may come between 98-102 days and will finished after 110 days.
- the peak on pandemic in state India, named**Kerala** may come between 85-90 days and will finished after 110 days.

References

1. https://www.who.int/news-room/q-a-detail/q-a-coronaviruses
2. https://www.kaggle.com/c/covid19-global-forecasting-week-4
3. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
4. https://www.kaggle.com/sudalairajkumar/covid19-in-india
5. https://towardsdatascience.com/time-series-machine-learning-regression-framework-9ea33929009a
6. https://towardsdatascience.com/global-covid-19-forecasting-with-linear-regression-and-arima-c154c163acc1

Acknowledgements

- World Health Organization (WHO)
- Johns Hopkins University for making the data available for educational and academic research purposes
- Indian Ministry of Health & Family Welfare for making the data available to general public.
- covid19india.org for making the individual level details and testing details available to general public.
- Wikipedia for population information.