**A**

**Project Report**

**on**

**Big Data Analytics**

**Breaking Down Covid-19**

**Submitted by-**

**Yarala Hruthik Reddy**

**1700289C203**

**CSE-1**

**Nikita Agarwala**

**1700xxxC203**

**CSE-1**

**Under Guidance of-**

**Dr. Yogesh Gupta**

Assistant Professor

******

**Department of Computer Science and Engineering**

**SCHOOL OF ENGINEERING AND TECHNOLOGY**

**BML MUNJAL UNIVERSITY GURGAON-122413, INDIA**

***DECEMBER 2020***

**Acknowledgement**

We are extremely grateful for a course that allows us to build projects and let us attain the inherent experience. We would like to thank our mentor Dr. Yogesh Gupta for enabling us to choose a project that solves real-world issues.

Thanking You

Nikita Agarwala

Yarala Hruthik Reddy

**INDEX**

[Abstract 4](#_Toc57920036)

[Motivation 5](#_Toc57920037)

[1. INTRODUCTION 6](#_Toc57920038)

[2. PROBLEM STATEMENT 8](#_Toc57920039)

[2.1 Global Analysis 8](#_Toc57920040)

[2.2 Forecast 8](#_Toc57920041)

[2.3 Indian Analysis 8](#_Toc57920042)

[3. LITERATURE REVIEW 8](#_Toc57920043)

[3.1 Existing State of Art 8](#_Toc57920044)

[3.2 Details of the Existing State of Art 8](#_Toc57920045)

[4. METHODOLOGY 9](#_Toc57920046)

[4.1 Technical Features 9](#_Toc57920047)

[4.2 Block Diagram 9](#_Toc57920048)

[4.3 Technologies Used 9](#_Toc57920049)

[5. RESULTS and DISCUSSION 9](#_Toc57920050)

[6. CONCLUSIONS and FUTURE WORK 9](#_Toc57920051)

[REFERENCES 9](#_Toc57920052)

# Abstract

# Motivation

# INTRODUCTION

In the family of Coronaviruses that have caused epidemics such as SARS (2002-2004), MERS (2012-Present) and COVID-19 (2019-Present), only COVID-19 has been declared a pandemic by the WHO. The number of confirmed cases crossing 60 million and reported deaths crossing 1.4 million as of 29th November 2020 as reported in the weekly epidemiological update of WHO [1]. Though not deadly as its counterparts SARS and MERS but is more contagious than both. Covid-19 and SARS have similar base R0,but the symptoms of SARS were more visual than Covid-19 resulting in less spread i.e., there was decent contrast between the healthy and the infected. Covid-19 was also noted to spread through asymptomatic patients [2].

Unlike MERS (Except for the ROK outbreak) and SARS, Covid-19 outbreak epicenters kept shifting. Initial epicenter was China, then shift was towards Europe (Italy and neighboring countries) and currently the US is the new epicenter of Covid-19 [3].

A picture containing text, screenshot, indoor

Description automatically generated

Figure 1. Almost every country has reported cases.

Containment measures included social distancing, contact tracing, quarantine of suspected cases and isolation of all contracted cases. The other coronaviruses outbreaks did not call for social distancing. But SARS-CoV-2 spreads faster, and it affects the old-aged people disproportionately. Social distancing helps keep further infected people from spreading the virus while keeping National Healthcare Systems from overwhelming.

\*Nikita write about India in similar fashion\*

# PROBLEM STATEMENT

Any patterns identified in the dynamics of Covid-19 can be crucial. These patterns could help devise strategies in better handling the pandemic. We are analyzing two datasets: a global dataset and an Indian dataset of cases. A forecast of the cases is also done for select countries using the Holt-Winter’s Model [4]. \*Nikita do something here\*

# LITERATURE REVIEW

## Existing State of Art

## Details of the Existing State of Art

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. | Existing state of art | Drawbacks in existing state of art | Overcome |
| 1 |  |  |  |

# METHODOLOGY

## Datasets

We had two datasets relating to Covid-19. A global dataset which had daily data of cases from every country that reported Covid-19 and an Indian dataset which had district wise data. The global dataset was obtained from Our World In Data (OWID) git repository [5] and the Indian dataset was obtained from a volunteer driven and crowd sourced database [6].

The OWID dataset had 52873 rows with 41 columns. \*Nikita continue here about your dataset, how you added those districts and all\*

## Global Analysis

## Exploratory Data Analysis

We did an exploratory data analysis on our dataset to narrow down the selection to few countries that were at the risk of Covid-19. The countries we chose based on this analysis are:

Table 1. Countries chosen for analysis.

|  |  |
| --- | --- |
| No | Country |
| 1 | United States |
| 2 | Sweden |
| 3 | Bolivia |
| 4 | Serbia |
| 5 | Japan |
| 6 | Canada |
| 7 | India |
| 8 | Ireland |
| 9 | Denmark |
| 10 | Mexico |
| 11 | Chile |
| 12 | Belgium |
| 13 | United Kingdom |
| 14 | France |
| 15 | Italy |

Figure 2 shows the scatter plot between total deaths per million and median age as of 27-10-2020. This figure gives insights into the effect of Covid-19 on countries with respect to median age.

Chart, scatter chart

Description automatically generated

Figure 2.

Countries with median age greater than 35 and high deaths:

Table 2. Countries with median age greater than 35 and high deaths.

|  |  |
| --- | --- |
| No | Country |
| 1 | Chile |
| 2 | United States |
| 3 | Belgium |
| 4 | United Kingdom |
| 5 | Sweden |
| 6 | France |
| 7 | Italy |

But there are a lot of other countries with similar and even greater median age but less deaths. The reason being the population greater than 65 was higher for the countries in table 2. This was noticed in figure 3 when a similar plot was made between total deaths per million and population aged more than 65 years.

Chart, scatter chart

Description automatically generated

Figure 3.

Countries with greater population of 65 aged people high deaths:

Table 3. Countries with greater population of 65 aged people high deaths.

|  |  |
| --- | --- |
| No | Country |
| 1 | Belgium |
| 2 | United States |
| 3 | United Kingdom |
| 4 | Sweden |
| 5 | France |
| 6 | Italy |

These countries in the table 3 were also the worst hit by Covid-19 when the death rates are compared to other countries. This means Covid-19 targets an inherently older population.

Yet Japan with the highest median age and the oldest population had significantly less deaths. This shows that Japan had an edge over the virus though being highly susceptible to it.

Figure 4 and figure 5 show the general situation of the USA and Japan, respectively. It is clear that the USA did not attempt enough to contain the virus while Japan came close to containing the virus before a new wave of infections started once again.

Chart, line chart, histogram

Description automatically generated

Figure 4. USA. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

A picture containing graphical user interface

Description automatically generated

Figure 5. Japan. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

## Efficiency of Lockdowns

We plotted new cases per million of Sweden and India who implemented the minimum and the maximum restrictions respectively to measure the efficiency of lockdowns had on the spread of the virus. Figure 6 and figure 7 show the respective plots.

Chart, line chart, histogram

Description automatically generated

Figure 6. Sweden. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

A picture containing chart

Description automatically generated

Figure 7. India. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

Early restrictions in India helped in containing the virus to a large extent even though being the second most populous country. Unlike Sweden, which saw frequent rises and falls in its new cases. Though no mandatory restrictions were imposed, the public were well aware of the virus and tool voluntary measures to curb the spread.

## Diabetes Effect on Death Rate

The effect of diabetes on Covid-19 was not much pronounced implicating it had negligible effect on the death rate. The comparison between the USA and Ireland gave this insight. The USA has the highest prevalence rate for diabetes while Ireland has the least. Figure 8 and figure 9 show the respective plots.

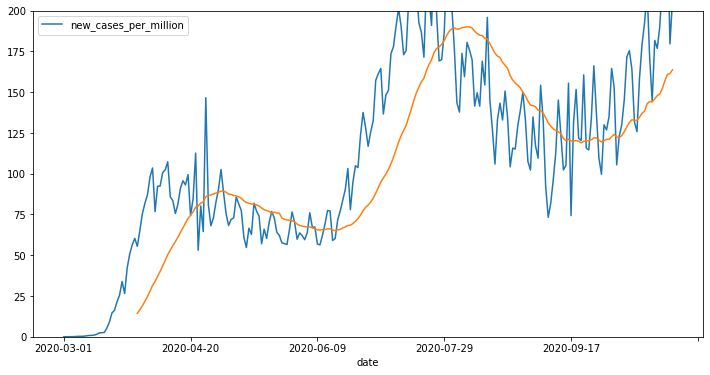


Figure 8. USA. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

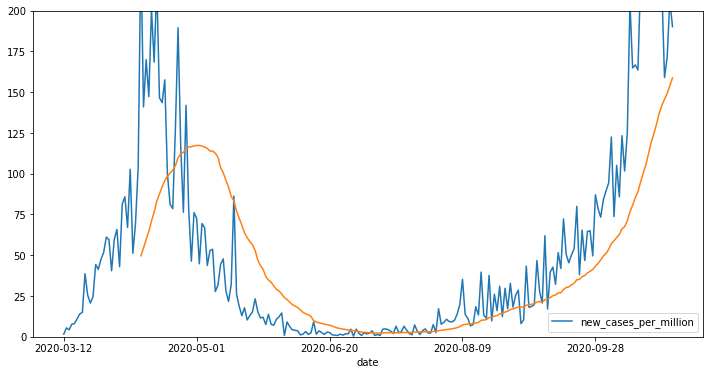


Figure 9. Ireland. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

## Obesity Effect on Death Rate

Unlike diabetes, obesity had a more pronounced effect on the death rate due to the virus. The comparisons between the USA, Canada, India, and Japan gave this insight. Figures 10, 11, 12 and 13 show the respective plots.

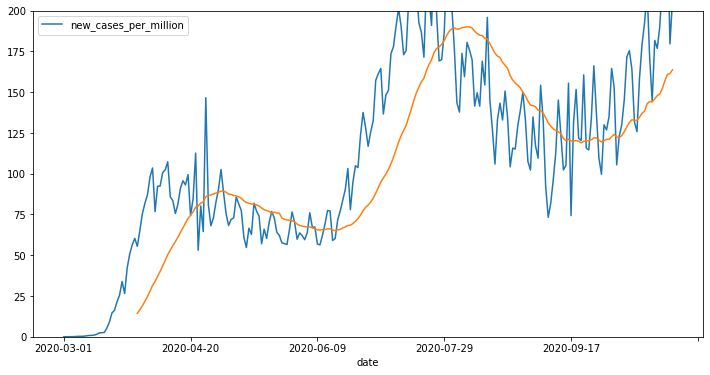


Figure 10. USA. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

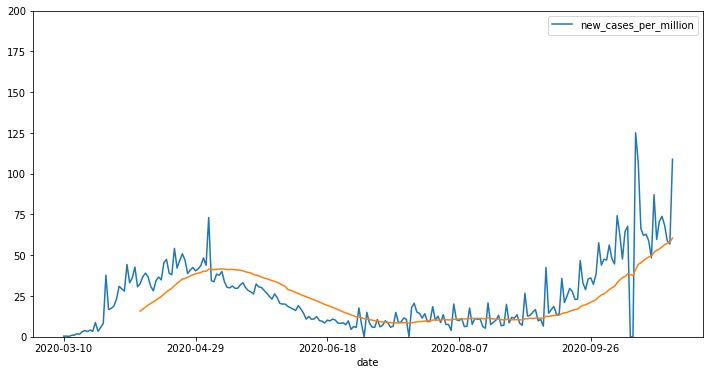


Figure 11. Canada. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

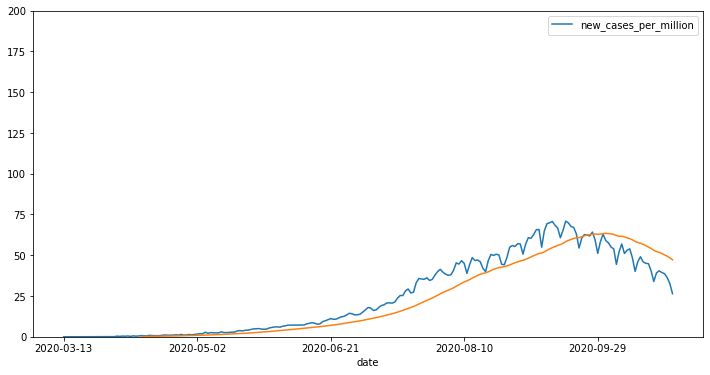


Figure 12. India. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

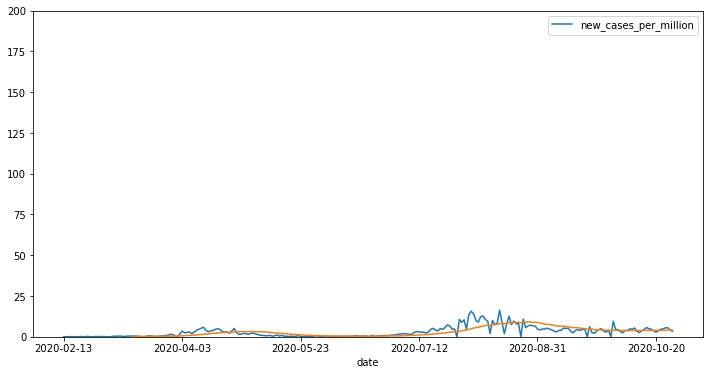


Figure 13. Japan. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

Another reason for Japan having less death rate could be due to its healthy population relative to other countries. Also, the reason for India even with a greater population having less death for a contagious virus.

## Testing Effect on Death Rate

The effect of testing was pronounced for certain countries such as Denmark and Mexico who have had the highest and the least testing capacity respectively. Figure 14 and figure 15 show the respective plots.

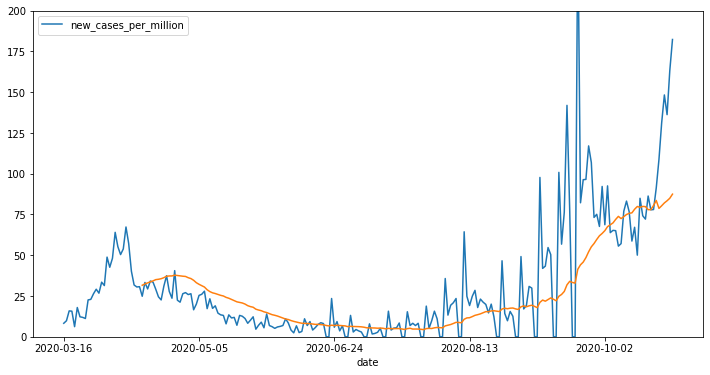


Figure 14. Denmark. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

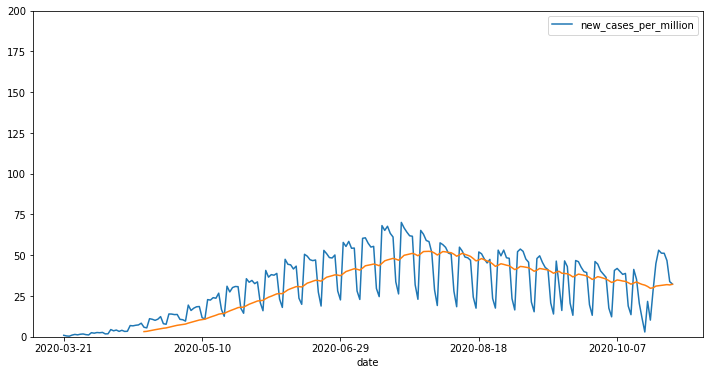


Figure 15. Mexico. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

## Forecast

We used the Holt-Winter’s Model which models three aspects of time series being the average, trend, and seasonality. In doing so it tries to predict the future provided the series is seasonal and repetitive over time.

The model was run on three countries: United States, Brazil, and Japan. There were different training and testing sets for each of these as the number of entries were different. The forecast was done 100 days into the future with additive model and a seasonal period of 14.

In order to test the forecast, we used data collected up to the month of October though we had data available till November. Our forecast showed that November month matched closely to the actual scenario of the month.

The accuracy of the model ranged from 90% to 95% depending on the country and the amount of data available. The accuracies for USA, Brazil, and Japan were 94%, 92%, and 93% respectively.

Chart, line chart

Description automatically generated

Figure 16. USA. The orange line indicates the forecast.

A picture containing chart

Description automatically generated

Figure 17. Brazil. The orange line indicates the forecast.

Chart

Description automatically generated

Figure 18. Japan. The orange line indicates the forecast.

## Indian Analysis

## Technologies Used

# RESULTS and DISCUSSION

# CONCLUSIONS and FUTURE WORK

# REFERENCES

|  |  |
| --- | --- |
| [1] | World Health Organisation, "Weekly epidemiological update - 1 December 2020," 1 December 2020. [Online]. Available: https://www.who.int/publications/m/item/weekly-epidemiological-update---1-december-2020. [Accessed 2 December 2020]. |
| [2] | Centers for Disease Control and Prevention, "Covid-19 FAQs," 6 May 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html. |
| [3] | J. Dubey, "Covid-19 Factoid: Shifting of epicentre to the US, and 3 other data trends," 1 April 2020. [Online]. Available: https://www.business-standard.com/article/current-affairs/covid-19-factoid-shifting-of-epicentre-to-the-us-and-3-other-data-trends-120040100168\_1.html. |
| [4] | R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," 2018. [Online]. Available: OTexts.com/fpp2. [Accessed 2 December 2020]. |
| [5] | D. B. D. G. C. G. J. H. B. M. E. M. E. O.-O. H. R. M. R. Cameron Appel, "Data on COVID-19 (coronavirus) by Our World in Data," [Online]. Available: https://github.com/owid/covid-19-data/tree/master/public/data. [Accessed 23 October 2020]. |
| [6] | COVID-19 India Org Data Operations Group, "{Dataset for tracking COVID-19 spread in India," [Online]. Available: https://github.com/covid19india/api. [Accessed 23 October 2020]. |