Covid19: Data Analytics and Prediction

Karan Patel
CMPE-256
San Jose State University
San Jose, USA
karan.patel@sjsu.edu

Ajith Balaji Nagarajan
CMPE-256
San Jose State University
San Jose, USA
ajithbalaji.nagarajan@sjsu.edu

Pooja Ramaswamy
CMPE-256
San Jose State University
San Jose, USA
pooja.ramaswamy@sjsu.edu

Roopesha Sheshappa Rai
CMPE-256
San Jose State University
San Jose, USA
roopeshasheshappa.rai@sjsu.edu

***Abstract*—The Coronavirus disease 2019 (COVID-19) is a respiratory illness caused by a new coronavirus that has shaken the world and disrupted normal lives. There is no specific treatment or vaccine that has been developed thus far, to treat this disease. This has developed increasing concern globally, leading to it being declared as a pandemic. In this project, the data obtained from John Hopkins University, which is updated on a daily basis will be used to identify the Key Performance Indicators (KPIs) to perform various data analytics and present the analytics in a format that could be used for making key decisions as well as serve as a medium for presenting the information obtained from these analytics in a storytelling format for establishing effective communication with data. Along with that, epidemiological models such as SIR (Susceptible, Infectious, Recovered) models along with several others will be used to evaluate and present a forecast of the future impact that could be caused by this disease [1].**

***Index Terms*—SIR model, Spark Data Analytics, COVID-19, Predictive Analytics**

I. Introduction

The outbreak of Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome also known as SARS or coronavirus-2 (SARS-CoV-2), has thus far killed about 283,023 people and infected over 4,165,617 all over the world, resulting in catastrophe for humans. Similar to its homologous virus SARS-CoV, which caused SARS in thousands of people in 2003, SARS CoV-2 might also be transmitted from bats and causes similar symptoms through a similar mechanism. The mortality rate analysis in figure 1.1 shows the comparison between the various epidemics that affected the world in the past. Ebola had the highest mortality rate of 39.53, followed by MERS with a mortality rate of 34.4. COVID-19 has a mortality rate of 6.54 which is less severe compared to EBOLA, SARS, and MERS as mentioned earlier but is more severe than H1N1 with a mortality rate of 0.29. Although, COVID-19 has lower severity and mortality than SARS but COVID-19 is much more transmissive and affects more elderly individuals than youth and more men than women. The outbreak of coronavirus disease 2019 (COVID-19) has created a global health crisis that has had a deep impact on the way we perceive our world and our everyday lives. Not only the rate of contagion and patterns of transmission threatens our sense of agency, but the safety measures put in place to contain the spread of the virus also require social distancing by refraining from doing what is inherently human, which is to find solace in the company of others. There is no specific treatment or vaccine that has been developed as of now to treat this disease and this has developed an increasing concern globally, leading to it being declared as a pandemic. The motive of this project can be seen in two parts. Firstly, to identify the Key Performance Indicators (KPIs) to perform various data analytics and present the analytics in a format that could be used for making key decisions as well as serve as a medium for presenting the information obtained from these analytics in a storytelling format to communicate effectively with data. Secondly, to evaluate and present a forecast of the future impact that could be caused by this disease using epidemiological models such as SIR models which stands for Susceptible, Infectious, Recovered along with several others.

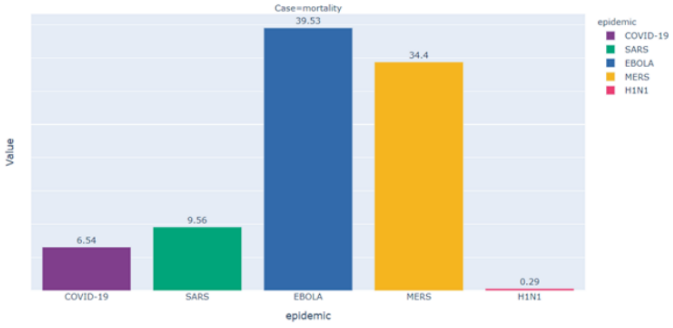

Fig. 1. Mortality rate comparison of different global pandemics

II. Literature Review

Liu et al investigated the H1N1 epidemic pattern in Hong Kong since the diagnosis of the first identified case there. He stressed on the fact to build a mathematically based epidemic model, SIR model to analyze and predict the pattern of infection and development of the H1N1 disease in Hong Kong. His work also stresses on the need to design these mathematical epidemiological models to provide the condition and prevention methods for the sanitation department. His mathematical analysis of the pattern and predictions aligned closely with the real-world data and hence valuable insights from this paper has been incorporated in the analysis performed within the scope of this paper [1].

Yang et al proposed the need to take the saturating contact rate of individual contacts during the calculation of the mathematical epidemiological model. Their work analyzed the

saturating incidence and the age of infection to build the dynamic behavior of the epidemiological model that could also determine the effectiveness of the prediction. Valuable insights from this paper have been well studied in formulating the model proposed in this paper [2].

Chen et al investigated the Ebola virus disease epidemic pattern, a disease that has been spread through animal bloods and secretion, could be easily related to the COVID-19 , which has also been believed to be spread from animals. Hence, valuable insights related to the analysis and prediction could be easily related to the analysis of COVID-19, as the analysis had been carried out in this paper including the analysis carried out during the initial stages of Ebola when there were no vaccines being developed [3].

Maki et al proposed a stochastic differential equations method to solve the SIR model equations for SARS pandemic in 2003. Such stochastic differential equations which provide valuable insights in economics and finance when applied to SIR models can provide invaluable insights, which could be used effectively to study the patterns and propose better methods to solve or prevent the disease. Since, COVID-19 is still in the early stages or what the experts call it to be wave 1 of the pandemic, this analysis has not been incorporated into this paper [4].

Piccini et al also proposed a stochastic based approach to solve the differential equations that govern the SIR models.They followed a graph based approach where the connected components of the graph represented the spread of disease along those connected through the edges. The deletion of a node in the graph was directly correlated to the immunization of that node to the disease. Such kind of advanced predictive analytics can be used in the future to find out the root cause of the disease and make better decisions for the future [5].

III. Data

This project analyzes live large data by querying datasets and putting it into the spark data frame, later these data frames will be processed again to get the desired data. Following APIs are developed as part of the project:

1) get_global_info() Get global level data. Returns spark dataframe with columns - 'NewConfirmed', 'NewDeaths', 'NewRecovered', 'TotalConfirmed', 'TotalDeaths', 'TotalRecovered'
2) get_countries_info() Get data for all countries. Returns spark dataframe with columns - 'Country', 'Date', 'NewConfirmed', 'NewDeaths', 'NewRecovered', 'TotalConfirmed', 'TotalDeaths', 'TotalRecovered'
3) get_top_countries_totalconfirmed(country_count=10) Get total confirmed cases for top most affected countries. Returns spark dataframe with columns - 'country', 'TotalConfirmed'
4) get_top_countries_totalrecovered(country_count=10) Get total recovered cases for top most countries. Returns spark dataframe with columns - 'country', 'TotalRecovered'
5) get_top_countries_recover_rate(country_count==10) Get the total recovered rate for top most countries. Returns spark dataframe with columns - 'country', 'RecoveredRate'
6) get_top_countries_deaths_rate(country_count=10) Get the total death rate for top most countries. Returns spark dataframe with columns - 'country', 'DeathsRate'
7) get_countries_less_recover_rate(country_count=10) Get the least recovered rate by top most countries. Returns spark dataframe with columns - 'country', 'RecoveredRate'
8) get_countries_less_deaths_rate(country_count=10) Get the least death rate by top most countries. Returns spark dataframe with columns - 'country', 'DeathsRate'
9) get_country_status(country) Get the country level data. Returns spark dataframe with columns - 'NewConfirmed', 'NewDeaths', 'NewRecovered', 'TotalConfirmed', 'TotalDeaths', 'TotalRecovered'
10) get_country_status(country, start_time, end_time) Get the country level data based on duration. Returns spark dataframe with columns - 'NewConfirmed', 'NewDeaths', 'NewRecovered', 'TotalConfirmed', 'TotalDeaths', 'TotalRecovered', 'StartDate', 'EndDate'
11) get_world_status_timebased(start_time, end_time) Get the global level data based on duration. Returns spark dataframe with columns - 'Country', 'NewConfirmed', 'NewDeaths', 'NewRecovered', 'TotalConfirmed', 'TotalDeaths', 'TotalRecovered', 'StartDate', 'EndDate'
12) get_all_status() Get complete details including global level summary and all countries details.

IV. Approach

The project has been broadly subdivided into two parts: Data Analytics and Prediction. The data analytics part uses spark real-time analytics to present various Key Parameter Indications (KPIs) that could be used more effectively in the decision making process as well to present the information to the user in a more easily interpretable manner. The second part involves using neural networks to present an estimation of the number of infected cases and deaths for the next 20 days and the SIR model infection curve to identify the peaks and an estimated amount of days for the curve to flatten.

A. Spark Process

Dealing with massive amounts of data often requires parallelization and cluster computing; Apache Spark is an industry standard for doing just that. In this lab we introduce the basics of PySpark, Spark's Python API, including data structures, syntax, and use cases as stated in [6].

Apache Spark is a general-purpose distributed large scale data processing engine and allows to process data with the help of stream processing, SQL, mlibs, and other modules. Apache Spark is written in Scala language but it has support for other programming languages like Python, Java, R along with Scala. This project obtains data from John Hopkins University in the form of COVID-19 REST APIs response and CSV file.

'Pyspark' is a Python module used to process these live data and convert data into Spark data frames.

B. Neural Network Model

Logistic regression and artificial neural networks are the models of choice in many medical data classification tasks. In this review, we summarize the differences and similarities of these models from a technical point of view, and compare them with other machine learning algorithms. We provide considerations useful for critically assessing the quality of the models and the results based on these models as stated in [7].

For predicting the confirmed cases and deaths we are using regression and for predicting it in a better way we are using Neural Network as our model will be complex. As suggested in [8] the most basic model for the regression of the Neural Network. We are using the Deep Neural Network Models where there is more than one hidden layer in our case we have 3 hidden layers. Neural Network for regression consists of Net Input Function Activation Function and Unit Step Function. The figure 2 shows that there are total 3 neurons one for input function one for activation function and one for unit step function there is also error which is after activation function and before unit step function and there is input before net input function and output after Unit Step Function.

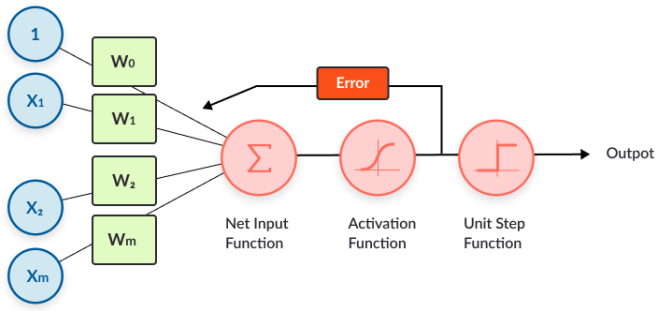

Fig. 2. Single Neuron Regression Model

We have used 3 hidden layers of size 80 for each layer we had used the dense layer and Leaky Relu after the Dense Layer. We had used Adam optimizer with learning rate of 0.001. The total parameters we have used are 13,201. We had fit the model in 1000 epoch.

C. SIR Model

SIR is one of the most famous epidemiological models.SIR is a simple model that considers a population that belongs to one of the following states:

1) **Susceptible (S)** The individual hasn't contracted the disease, but she can be infected due to transmission from infected people
2) **Infected (I)** This person has contracted the disease
3) **Recovered/Deceased (R)** The disease may lead to one of two destinies: either the person survives, hence developing immunity to the disease or the person is deceased.

There are many versions of this model, considering birth and death (SIRD with demography), with intermediate states, etc. However, since we are in the early stages of the COVID-19 expansion and our interest is focused in the short term, we will consider that people develop immunity (in the long term, immunity may be lost and the COVID-19 may come back within a certain seasonality like the common flu) and there is no transition from recovered to the remaining two states.

With this, the differential equations that govern the system are:

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{1}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

where N = S + I + R will be total population under study.

The SIR model can be implemented in many ways: from the differential equations governing the system, within a mean-field approximation or running the dynamics in a social network (graph). For the sake of simplicity, a simple numerical method known as the Runge-Kutta method had been used in this paper to solve the three differential equations mentioned above, which govern the SIR epidemiological model.

V. Result

The result comprises the results of the various data analytics performed using the real-time COVID-19 data, presented in the form of data visualizations, with each describing the various KPIs that could aid in the decision-making process as well as provide useful information. The visualization that describes the various analytics obtained from the prediction algorithms is also included in this section.

A. Analytics

The analysis shown in the figure below shows the total cases of active confirmed cases, recovered cases and deaths due to COVID-19 cases all around the world so far over the past months. We can infer from this that out of the total confirmed cases that are above 3.5 Million roughly over 1 Million have recovered and the deaths are nearing 1.5 Million at the time this analysis result was taken and shown in 3.

1) Confirmed cases on Map of all countries: In figure 4 one can observe the countries or parts of the world which are highly affected by the pandemic. We observe that out of all the regions around the globe the highest number of cases is in the North America region.

2) Total confirmed cases and deaths reported: In figure 5 it illustrates the overall confirmed cases and deaths as of 5th May 2020 when these results were taken.

3) Daily new cases and countries affected over time: In figure 6 it shows the number of new cases everyday as of May 5 2020 and the number of new countries getting affected everyday.

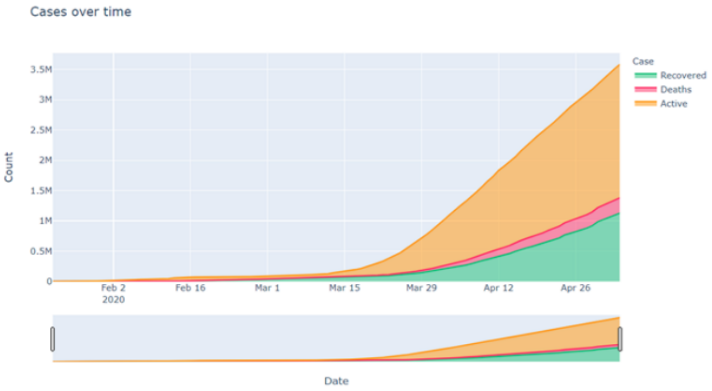

Fig. 3. Confirm, Death, Recovered of all Countries

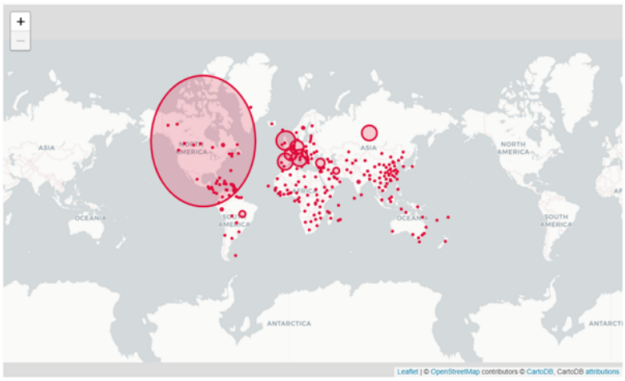

Fig. 4. Confirm Cases on Map of all Countries

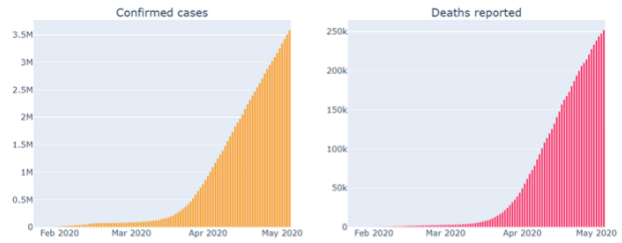

Fig. 5. Overall Confirmed Cases and Deaths

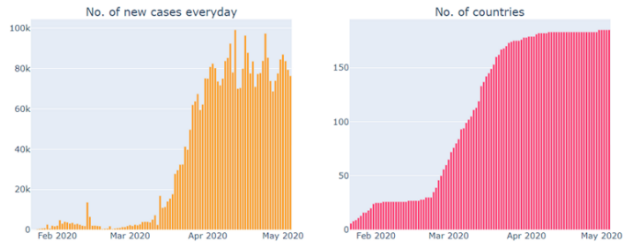

Fig. 6. New Countries affected by COVID19

4) Confirmed Cases of Top 15 Countries: In figure 7 it shows the top 15 countries with the most confirmed cases as of May 5,2020

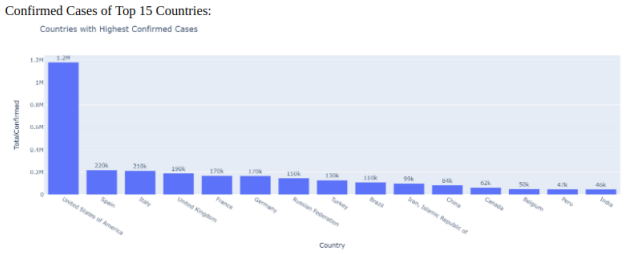

Fig. 7. Top 15 countries with highest confirmed cases

5) Deaths of Top 15 Countries: In figure 8 it shows the top 15 countries with the most deaths as of May 5,2020.

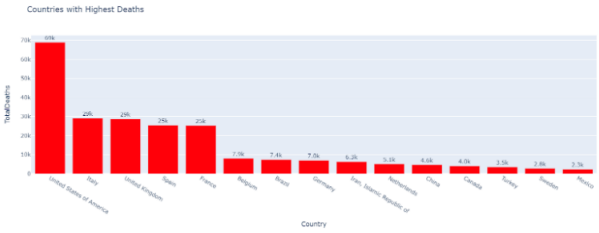

Fig. 8. Top 15 countries highest with Deaths

6) Recovered Cases of Top 15 Countries: In figure 9 it shows the top 15 countries with the most recovered cases as of May 5,2020.

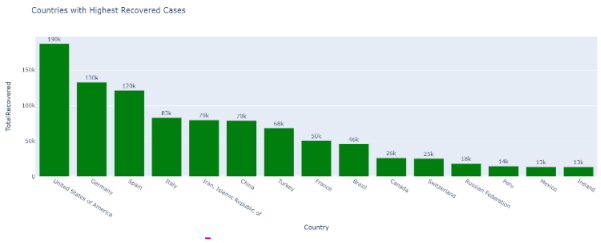

Fig. 9. Top 15 countries with highest Recovered Cases

7) Countries with highest Recovery Ratio: In figure 10 it shows the top countries with the highest case recovery ratio as of May 5,2020.

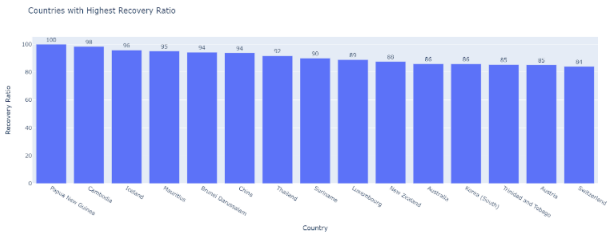

Fig. 10. Top 15 countries with highest Recovery Ratio

8) Countries with lowest recovery rate: In figure 11 it shows the countries with the lowest case recovery rate as of May 5,2020.

Table

	Country	RecoverRate	TotalRecovered	TotalConfirmed
0	Comoros	0.000000	0	3
1	United Kingdom	0.000000	0	182260
2	Papua New Guinea	0.000000	0	8
3	South Sudan	0.000000	0	45
4	Netherlands	0.000000	0	40236
5	Tajikistan	0.000000	0	76
6	Norway	0.004098	32	7809
7	Bangladesh	0.020137	177	8790
8	Equatorial Guinea	0.028571	9	315
9	Maldives	0.034682	18	519

Plot

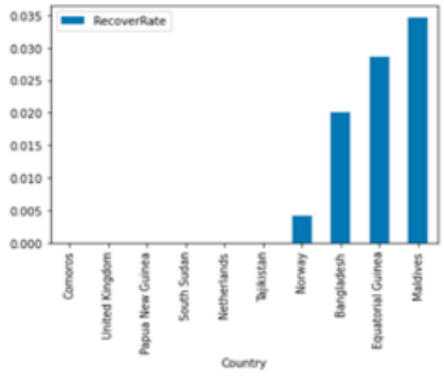

Fig. 11. Countries with lowest recovery ratio

9) Countries having higher death and recovery rate: In figure 12 it shows the countries having the highest death and recovery rate per 100 confirmed cases as of May 5,2020.

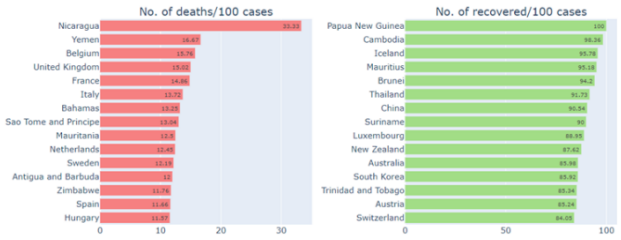

Fig. 12. Countries with higher death and recovery rate

10) Death vs confirmed cases log plot: In figure 13 it shows the log10 plot of the deaths vs confirmed cases. One can infer from this plot the countries that is most affected and less affected by the pandemic. It can be observed that compared to all the countries the United States is the most affected and has highest confirmed cases as well as deaths, followed by European countries such as the United Kingdom , Spain and France.

11) N days from 1000 case: In figure 14 below shows the variation of the number of cases since the country had first recorded its 1000th case. The plots show the confirmed cases that were reported after the 1000th case for every country. The flatter curve shows that the infection rate has decreased in those countries and an exponentially increasing curve shows that the country is still having an increase in the infection rate. It can be observed that the United States has a curve that is increasing exponentially, which shows that the confirmed cases are still increasing. However, in other countries it can

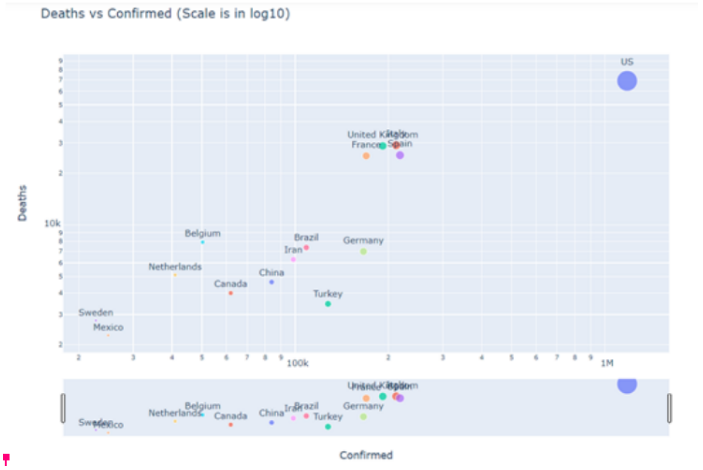

Fig. 13. Log plot of deaths vs confirmed cases

be observed that the curves are flattening showing that the number confirmed cases are becoming lesser or are constant indicating that these countries are recovering.

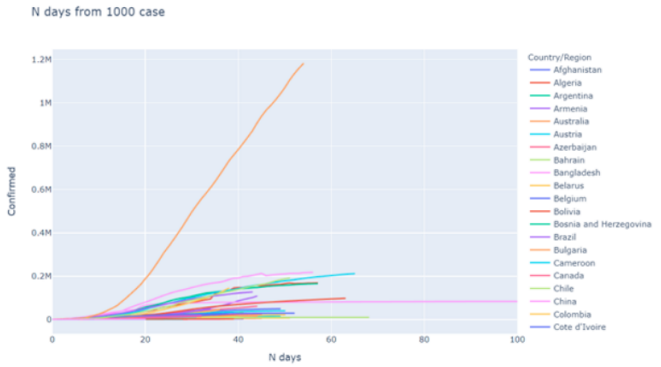

Fig. 14. Days from minimum 1000 cases

B. Prediction

We had done two prediction the first is with the help of Neural Network and second is with the help of SIR Model. In first we had predicted the graph for coming 20 days and we had predicted both confirmed cases and deaths and for SIR model we had seen the confirmed cases trend for USA.

1) Neural Networks: The Neural Network proposed had estimated the confirm cases upto 4.5 M on 25^{th} May 2020 which is an increase of 1 M cases in coming 20 days which you can see in figure 15. We are seeing the increasing trend for the Confirmed Cases. It has also proposed 375K Deaths on 25^{th} May 2020 which is 125K Deaths in next 20 Days which we can see in the figure 16.

2) SIR Model: The SIR model proposed, had been used to estimate the model of the Infected case for the population of the United States. Since we are in the early stage of COVID-19 and for the sake of simplicity, the model had been built with the assumption that those recovered will not enter the other two stages of the SIR. It can be seen from the figure

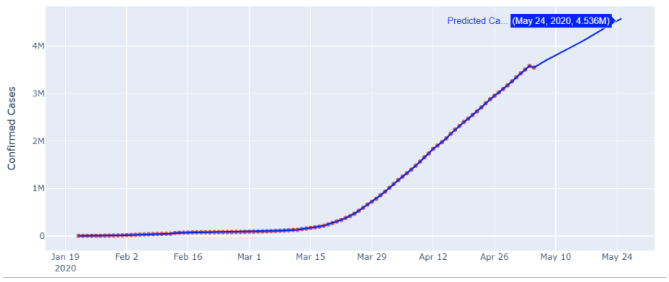

Fig. 15. Predicted of Confirmed Cases for next 20 days(Worldwide)

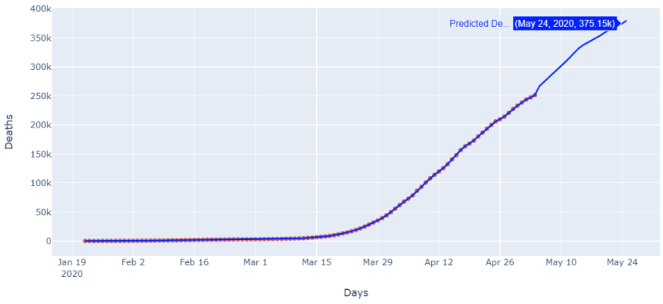

Fig. 16. Predicted of Deaths for next 20 days(Worldwide)

(number) that the number of estimated infected cases over time represents a bell curve and from the recent records it can be easily interpreted that the United States is currently in the declining part of the bell curve and the curve is anticipated to flatten out in around 30-40 days if the trend continues.

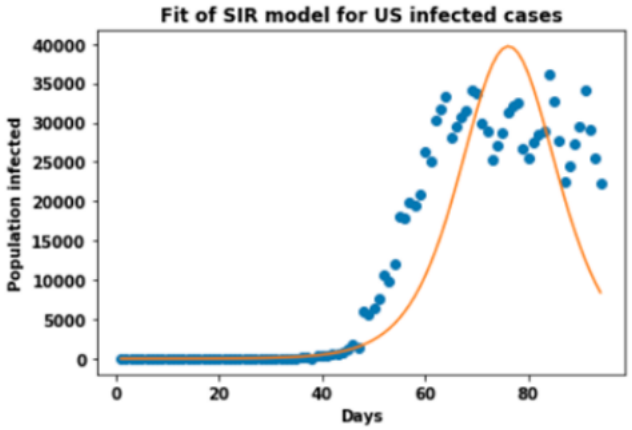

Fig. 17. Fit of SIR Model of Confirmed Cases of USA

VI. Conclusion

This paper discusses the various data analytics performed using spark describing various Key Parameter Indicators, which could aid in better decision making as well as provide useful information about the disease, which is still in the early stages. The predictive analytics using neural networks and SIR models also gives a good perspective of what the future impact of this disease would be so that better plans and decisions can be implemented to control it. As mentioned earlier, since the disease in the earlier stages, the SIR model has been built without considering the vital dynamics, i.e, it is built based on the assumption that those recovered will become immune and will not enter the other two stages again. However, with time and with additional data, this model could be extended in the future to include the vital dynamics to provide a much better analysis of the disease.

References

[1] Yulian Liu. Investigation of prediction and establishment of sir model for h1n1 epidemic disease. In *2010 4th International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4. IEEE, 2010.

[2] Junyuan Yang, Fengqin Zhang, and Xiaoyan Wang. A class of sir epidemic model with saturation incidence and age of infection. In *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)*, volume 1, pages 146–149. IEEE, 2007.

[3] Wenzhi Chen. A mathematical model of ebola virus based on sir model. In *2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration*, pages 213–216. IEEE, 2015.

[4] Yoshihiro Maki and Hideo Hirose. Infectious disease spread analysis using stochastic differential equations for sir model. In *2013 4th International Conference on Intelligent Systems, Modelling and Simulation*, pages 152–156. IEEE, 2013.

[5] Juan Piccini, Franco Robledo, and Pablo Romero. Analysis and complexity of pandemics. In *2016 8th International Workshop on Resilient Networks Design and Modeling (RNDM)*, pages 224–230. IEEE, 2016.

[6] Apache Spark. Apache spark. *Retrieved January*, 17:2018, 2018.

[7] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.

[8] Donald F Specht et al. A general regression neural network. *IEEE transactions on neural networks*, 2(6):568–576, 1991.