International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
© Global Society of Scientific Research and Researchers
http://ijcjournal.org/

COVID-19 Outbreak Data Analysis and Prediction Modeling Using Data Mining Technique

Tajebe Tsega Mengistie*

NITM, Bijni Complex, Laitumkhrah, Shillong, Meghalaya 793003, India
Email: tsegataju2017@gmail.com, t19cs011@nitm.ac.in

Abstract

Nowadays, sustainable development is considered a key concept and solution in creating a promising and prosperous future for human societies. Nevertheless, there are some predicted and unpredicted problems that epidemic diseases are real and complex problems. Hence, in this research work, a serious challenge in the sustainable development process was investigated using the classification of confirmed cases of COVID-19 (new version of Coronavirus) as one of the epidemic diseases. Hence, the data mining predictive modeling method of data handling and predictive or forecasting the spread of COVID-19 virus. This research work mainly works on predicting or forecasting by using fbprophet. Prophet it is a python library package used for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonally, plus holiday's effect. It works best with time series that have a strong seasonal effect and several seasons of historical data. The model helps to interpret patterns of public sentiment on disseminating related health information and assess the political and economic influence of the spread of the virus.

Keywords: Prediction modeling; Analysis and Visualization; Time Series; Fbprophet; COVID-19.

1. Introduction

On the Human life different challenges happened, like war, virus and other dangerous diseases and to influence in our daily life. By this case a lots of people are passed away. Among those diseases or virus is novel coronavirus, this virus is happened in the central Chinese city of Wuhan in late December are being reported daily around the world. Start from December till April been affected 209 countries or territories.

----

* Corresponding author.

The virus had affected 1,771,514 people overall the world and the number of deaths had totaled 108,503 and also more than 280,000 people have recovered to date. One of the things that The COVID-19 outbreak is increasingly revealing is how pervasive the surveillance mechanisms developed in the last decade or so, have become. In an effort to contain the spread of the virus, governments all over the world are adopting various surveillance and monitoring technologies: tracking those who have been tested positive and informing the public about their monitoring the movements of individuals to ensure their compliance with the policies of quarantine or confinement (as for instance, in China, Israel, and Singapore, as well as in Italy, Germany and Austria and now the USA); or using such technologies to predict or forecast the impacts of this virus overall the world. So, this paper proposed Data mining predictive and analysis modeling and this predictive modeling or forecasting the process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex neural network, mapped out by sophisticated software. As additional data becomes available, the statistical analysis model is validated or revised. In the case of data mining, there are different predictive modeling analyses are there like decision tree, logistic regression, and time series analysis. Time series analysis comprises methods for analyzing time-series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values by using fbprophet. Prophet is a python library package used for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonally, plus holiday’s effect. It works best with time series that have a strong seasonal effect and several seasons of historical data. The model helps to interpret patterns of public sentiment on disseminating related health information and assess the political and economic influence of the spread of the virus.

1.1. Literature Review

In data mining milieu, disease prediction such as diabetes, heart diseases, and cancer prediction plays an important role. This method to design a software tool based on data mining techniques which is useful in the medical field. In this paper, the author addresses eye problems affecting people with diabetes and related disease and virus [1]. In this research study, a serious challenge of sustainable development was investigated using the regression analysis. According to the results, the regression algorithm has an appropriate performance to predict and classify the parameters of a case study affected by COVID-19, and the accuracies based on Wuhan datasets were equal to 95.7% and 85.7% for training and testing, respectively[2]. SEIR refers to Susceptible, Exposed, Infectious, and Removed or Recovered, respectively. It is based on the SIR model but adds the Exposed compartment as a variable. Susceptible refers to individuals who can catch the infection and may become hosts if exposed, Exposed are individuals who are already infected but are asymptomatic, Infectious are individuals who are showing signs of infection and can transmit the virus, Removed or Recovered are individuals who are previously infected but are no longer infectious and already immune to the virus [3]. COVID-19 is the new virus due to this no more research is done before so, I cannot list or mansion related works but those researchers they can’t not correctly predicted rather than explain graphically show the spread of this virus. This the paper clearly defines and predicted the recovered, death and confirmed cases to compare the current status of the COVID-19 virus and what I have predicted.

1.2. Proposed System

COVID-19 current problem statement described in the previous section, I propose a prediction and analysis model to predict the Outbreak of COVID-19 on the overall world on the three basic things. Those are recovery, confirmed and death cases using data mining techniques for the coming 10 days. For the prediction purposed I have proposed to the time series prediction data mining algorithm to forecast for the coming 10 days and analysis everything graphically and compare and contrast the current status of the COVID-19 and predicted values. To forecast the outbreak of the COVID-19 virus I will use the Fbprophet open-source python software which is developed by Facebook data scientists. Facebook data scientists purposely they have developed for such types of unseasonal disease or virus based on the given dataset to predict or forecast for the future and show the clearly mention the impacts.

2. Methodology

2.1. Data Mining Techniques

Data mining plays an important role in various fields such as artificial intelligence, machine learning, and database systems. Data mining also used in the medical field for mining of healthcare systems that help to discover hidden patterns and is very useful for disease prediction [4]. Data mining is the technique in which useful information is extracted from the raw data. The data mining is applied to accomplish various tasks like clustering, prediction analysis and association rule generation with the help of various Data Mining Tools and Techniques. In the approaches of data mining, clustering is the most efficient technique which can be applied to extract useful information from the raw data. The clustering is the technique in which similar and dissimilar types of data can be clustered to analyze useful information from the dataset. The clustering is of many types like density-based clustering, hierarchical clustering, and partitioning based clustering. Data-mining capabilities in Analysis Services open the door to a new world of analysis and trend prediction. These research works mainly focus on the design and implementation of a COVID-19 prediction the confirmed, recovered and death cases for the coming 10 days on the whole world. The predictive data-mining model predicts the future outcomes based on past records present in the d with known answers. Data mining will help figure out the future credit risk of the applicant and predict future credit history of the applicant by using past data. Classification is known as the procedure used to locate a model that best suits identified data sets or ideas. The model helps predict the class of objects when class labels are not available [5]. The most widely used for data mining prediction is time series forecasting methods i. It is also one of the most popular models in traditional time series forecasting and is often used as a benchmark model for comparison with any other forecasting method. It is often difficult to identify a forecasting model because the underlying laws may not be clearly understood. In addition, hydrological time series may display signs of seasonality and nonlinearity which traditional linear forecasting techniques are ill equipped to handle, often producing unsatisfactory results [6].

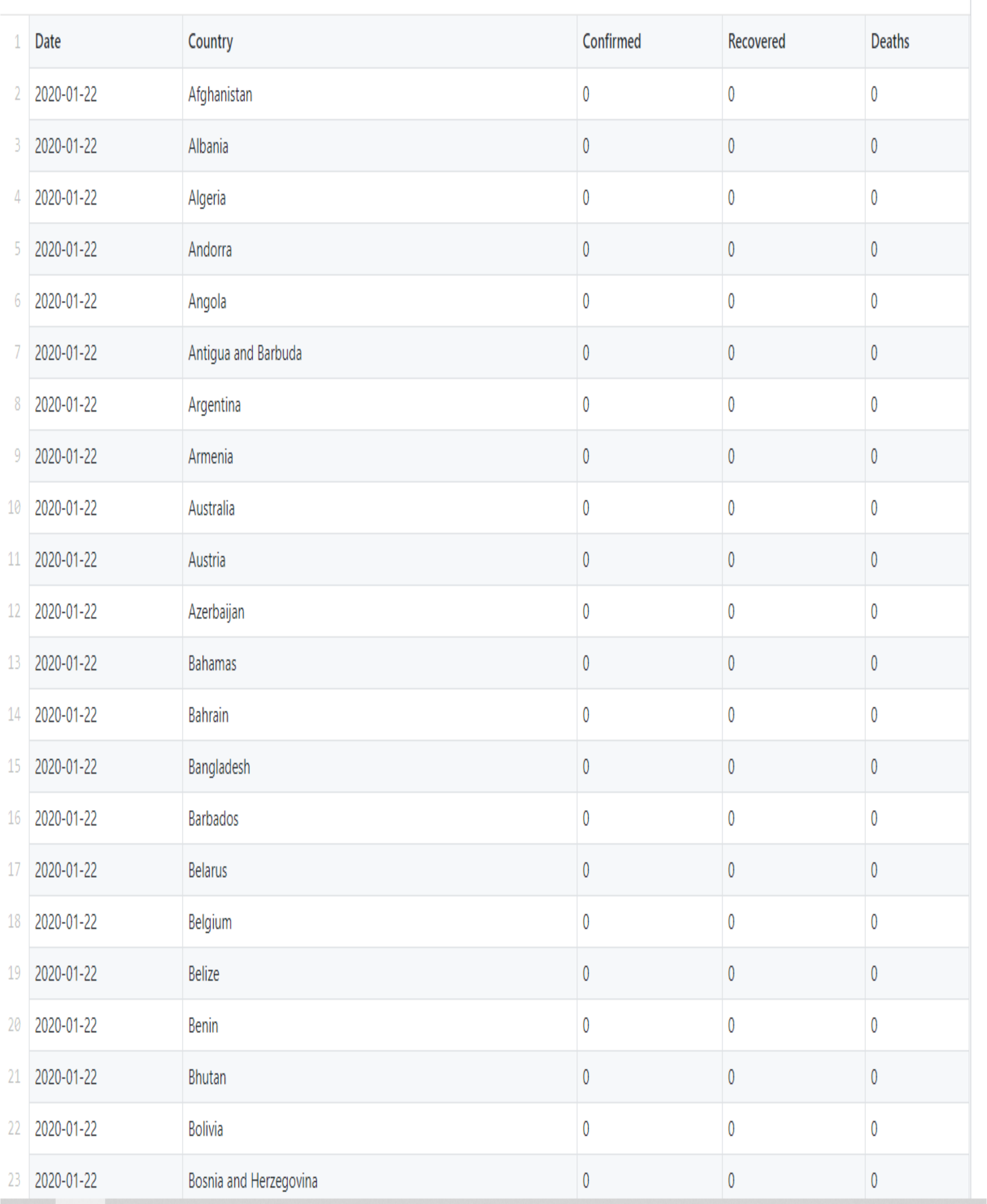
2.1.1. Time-series data mining

A time series is a sequence of data points recorded at specific time points - most often in regular time intervals (seconds, hours, days, months, etc.). Every organization generates a high volume of data every single day – be it sales figures, revenue, traffic, or operating cost. Time series data mining can generate valuable information for long-term business decisions, yet they are underutilized in most organizations. Below is a list of few possible ways to take advantage of time series datasets: Trend analysis: Just plotting data against time can generate very powerful insights. One very basic use of time-series data just understands the temporal pattern/trend in what is being measured. In businesses, it can even give an early indication of the overall direction of a typical business cycle. Outlier/anomaly detection: An outlier in a temporal dataset represents an anomaly. Whether desired (e.g. profit margin) or not (e.g. cost), outliers detected in a dataset can help prevent unintended consequences. Examining shocks/unexpected variation: Time-series data can identify variations (expected or unexpected) and abnormalities, detect signals in the noise. Association analysis: By plotting bivariate/multivariate temporal data it is easy (just visually) to identify associations between any two features (e.g. profit vs sales). This association may or may not imply causation, but this is a good starting point in selecting input features that impact output variables in more advanced statistical analysis. Forecasting: Forecasting future values using historical data is a common methodological approach – from simple extrapolation to sophisticated stochastic methods. Predictive analytics: Advanced statistical analysis such as panel data models (fixed and random effects models) rely heavily on multi-variant longitudinal datasets. These types of analysis help in business forecasts, identify explanatory variables, or simply help understand associations between features in a dataset. In my context, this predictive analysis is to predicting the COVID-19 outbreak and shows the gap or impact the whole world.

2.2. Data Source

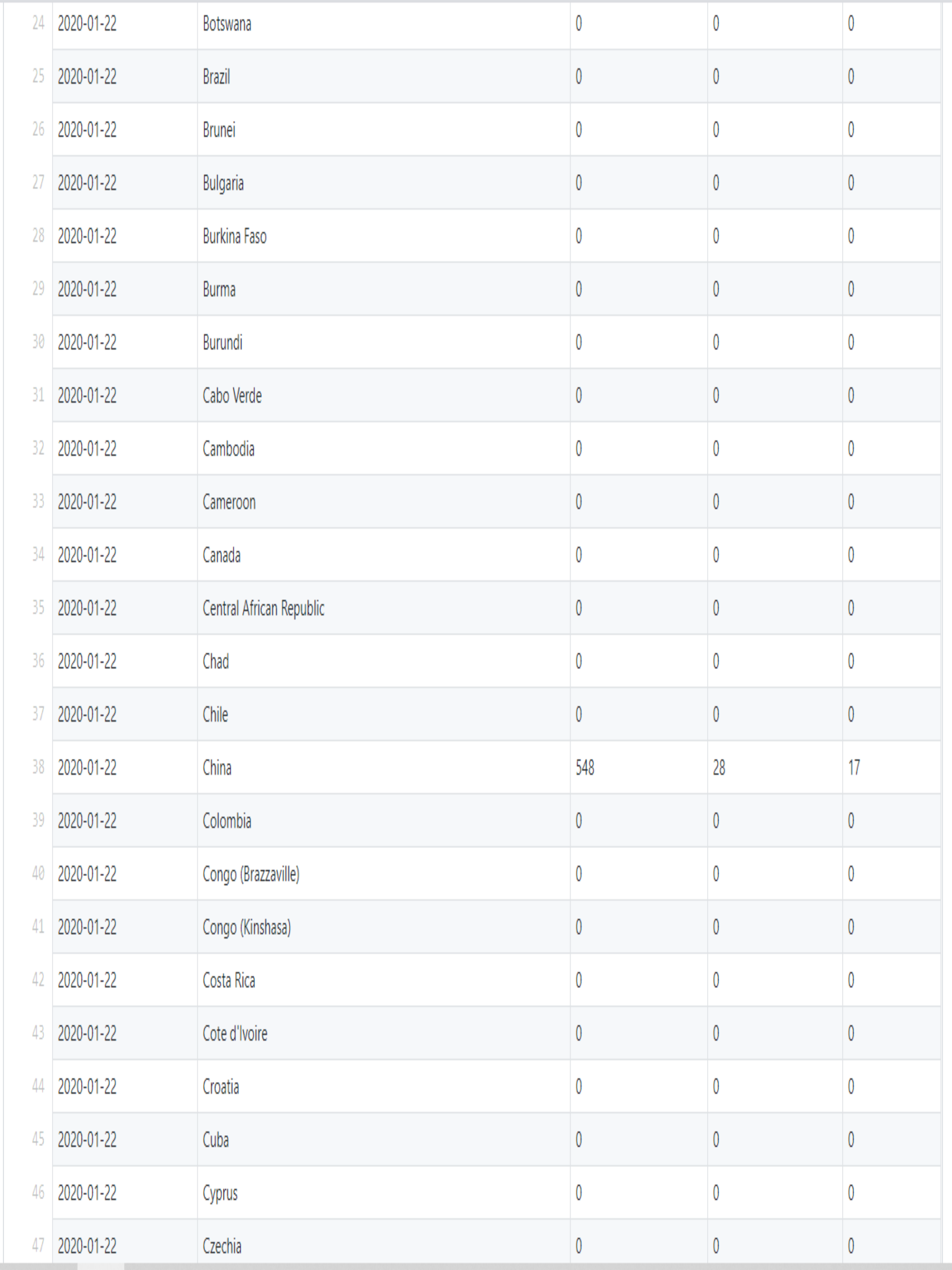
Data is extracted from verified sources such as John Hopkins University [4], WHO, Worldometr also GitHub.com, and DingXiangYuan, a website authorized by the Chinese government. The sites reported confirmed COVID-19 cases, as well as recovered and deaths for affected countries and regions. The COVID-19 module allows importing and displaying of data related to the 2019 Novel Coronavirus COVID-19 (2019-nCoV) from multiple sources. The intention is not only to make it easier to display and update data, but to select which data source to use depending on the information being displayed. For example, one data source would be used when displaying country-level data, but another, perhaps more current or accurate data source would be used when displaying data for a specific state, province, or county. There are a lot of official and unofficial data sources on the web providing COVID-19 related data. One of the most widely used dataset today is the one provided by the John Hopkins University’s Center for Systems Science and Engineering (JHU CSSE). Here is the Github link for the same: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE

Table 1: Reported cases as of 22 January 2020



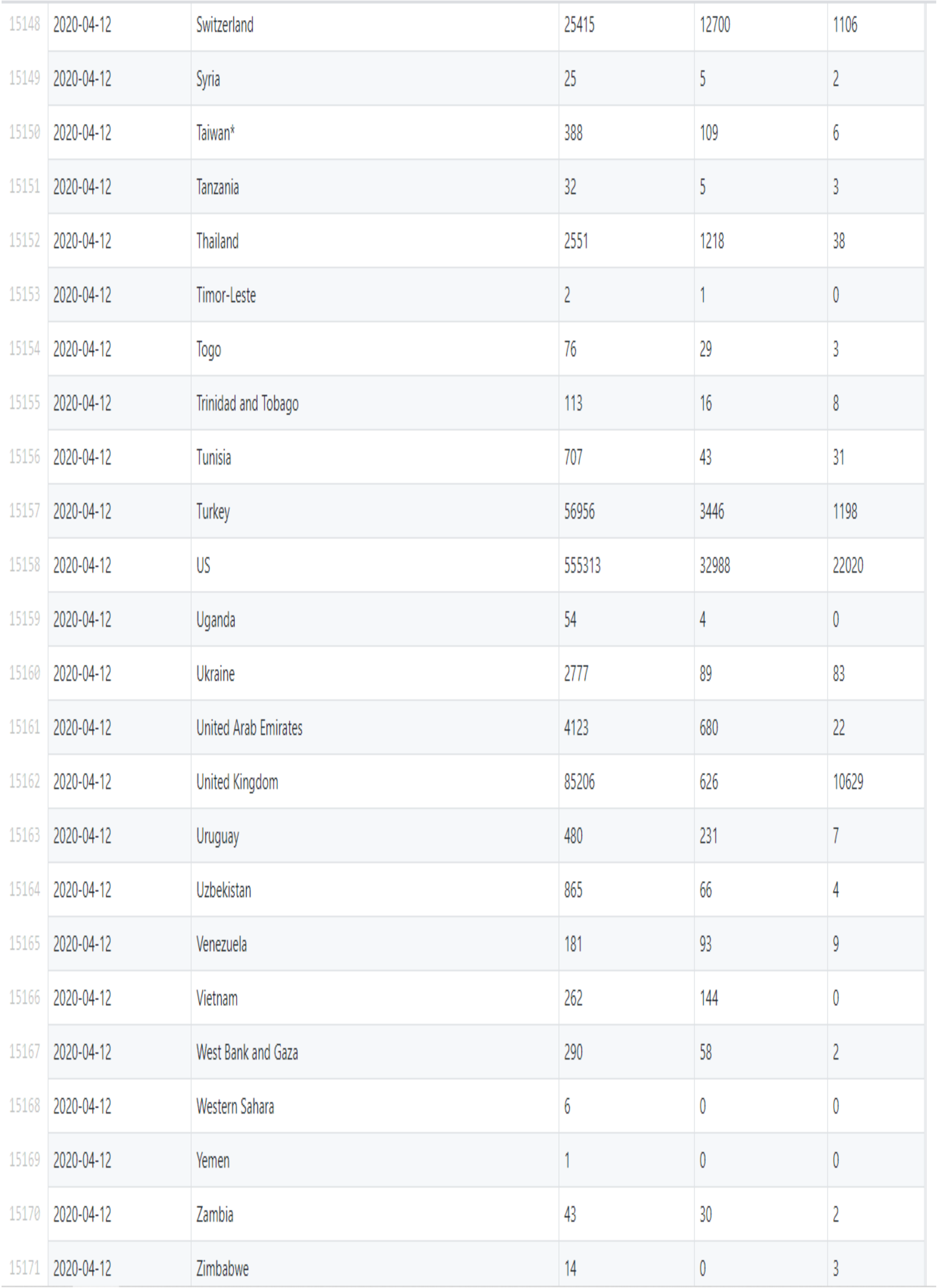
1	Date	Country	Confirmed	Recovered	Deaths
2	2020-01-22	Afghanistan	0	0	0
3	2020-01-22	Albania	0	0	0
4	2020-01-22	Algeria	0	0	0
5	2020-01-22	Andorra	0	0	0
6	2020-01-22	Angola	0	0	0
7	2020-01-22	Antigua and Barbuda	0	0	0
8	2020-01-22	Argentina	0	0	0
9	2020-01-22	Armenia	0	0	0
10	2020-01-22	Australia	0	0	0
11	2020-01-22	Austria	0	0	0
12	2020-01-22	Azerbaijan	0	0	0
13	2020-01-22	Bahamas	0	0	0
14	2020-01-22	Bahrain	0	0	0
15	2020-01-22	Bangladesh	0	0	0
16	2020-01-22	Barbados	0	0	0
17	2020-01-22	Belarus	0	0	0
18	2020-01-22	Belgium	0	0	0
19	2020-01-22	Belize	0	0	0
20	2020-01-22	Benin	0	0	0
21	2020-01-22	Bhutan	0	0	0
22	2020-01-22	Bolivia	0	0	0
23	2020-01-22	Bosnia and Herzegovina	0	0	0

Table 2: Reported cases as of 22 January 2020



24	2020-01-22	Botswana	0	0	0
25	2020-01-22	Brazil	0	0	0
26	2020-01-22	Brunei	0	0	0
27	2020-01-22	Bulgaria	0	0	0
28	2020-01-22	Burkina Faso	0	0	0
29	2020-01-22	Burma	0	0	0
30	2020-01-22	Burundi	0	0	0
31	2020-01-22	Cabo Verde	0	0	0
32	2020-01-22	Cambodia	0	0	0
33	2020-01-22	Cameroon	0	0	0
34	2020-01-22	Canada	0	0	0
35	2020-01-22	Central African Republic	0	0	0
36	2020-01-22	Chad	0	0	0
37	2020-01-22	Chile	0	0	0
38	2020-01-22	China	548	28	17
39	2020-01-22	Colombia	0	0	0
40	2020-01-22	Congo (Brazzaville)	0	0	0
41	2020-01-22	Congo (Kinshasa)	0	0	0
42	2020-01-22	Costa Rica	0	0	0
43	2020-01-22	Cote d'Ivoire	0	0	0
44	2020-01-22	Croatia	0	0	0
45	2020-01-22	Cuba	0	0	0
46	2020-01-22	Cyprus	0	0	0
47	2020-01-22	Czechia	0	0	0

Table 3: Reported cases as of 22 April 2020



15148	2020-04-12	Switzerland	25415	12700	1106
15149	2020-04-12	Syria	25	5	2
15150	2020-04-12	Taiwan*	388	109	6
15151	2020-04-12	Tanzania	32	5	3
15152	2020-04-12	Thailand	2551	1218	38
15153	2020-04-12	Timor-Leste	2	1	0
15154	2020-04-12	Togo	76	29	3
15155	2020-04-12	Trinidad and Tobago	113	16	8
15156	2020-04-12	Tunisia	707	43	31
15157	2020-04-12	Turkey	56956	3446	1198
15158	2020-04-12	US	555313	32988	22020
15159	2020-04-12	Uganda	54	4	0
15160	2020-04-12	Ukraine	2777	89	83
15161	2020-04-12	United Arab Emirates	4123	680	22
15162	2020-04-12	United Kingdom	85206	626	10629
15163	2020-04-12	Uruguay	480	231	7
15164	2020-04-12	Uzbekistan	865	66	4
15165	2020-04-12	Venezuela	181	93	9
15166	2020-04-12	Vietnam	262	144	0
15167	2020-04-12	West Bank and Gaza	290	58	2
15168	2020-04-12	Western Sahara	6	0	0
15169	2020-04-12	Yemen	1	0	0
15170	2020-04-12	Zambia	43	30	2
15171	2020-04-12	Zimbabwe	14	0	3

2.3. Data Visualization

Data Mining is used to find patterns, anomalies, and correlation in the large dataset to make the predictions using broad range of techniques, this extracted information is used by the organization to increase there revenue, cost-cutting reducing risk, improving customer relationship, etc. whereas data visualization is the graphical representation of the data and information extracted from data mining using the visual elements like graph, chart, and maps, data visualization tool, and techniques helps in analyzing massive amount of information and make decision on top of it. Machine learning makes it easier to conduct analyses such as predictive analysis, which can then serve as helpful visualizations to present. But data visualization is not only important for data scientists and data analysts, it is necessary to understand data visualization in any career.

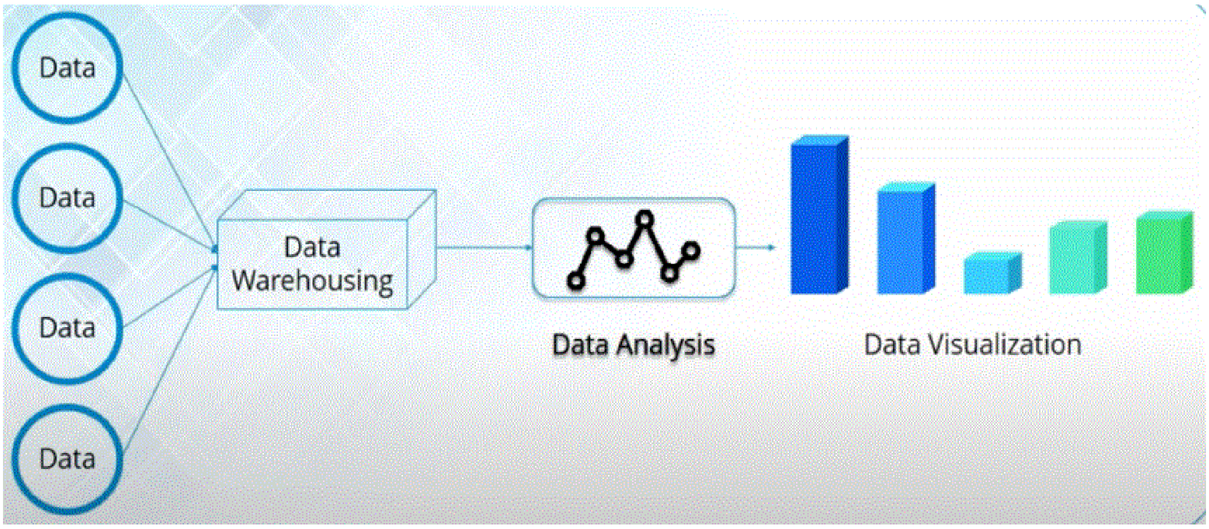
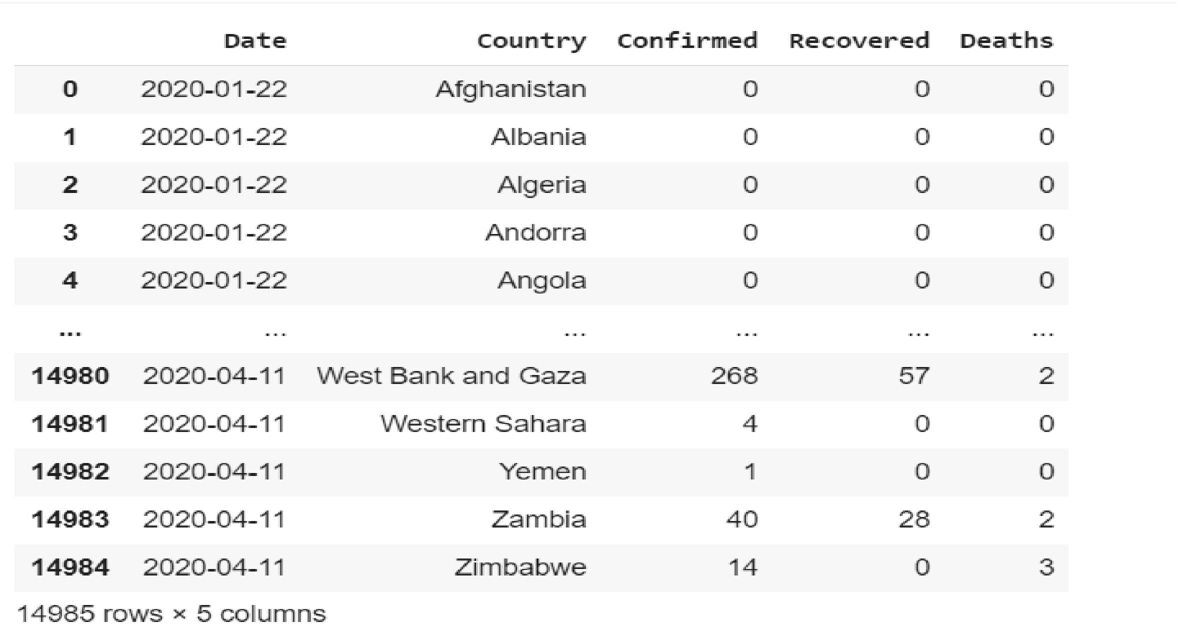


Figure 1: High level diagram ingestion for data visualization and analysis

2.4. Data Analysis

2.4.1. Analysis of COVID-19 cases in India

The data collected from different data sources especially the above mention data sources will be stored in the data warehouse. Then the stored data is preprocessed and analyzed by using the data mining modeling techniques and visualize it based on the given dataset. I have prepared dataset form data sources and this data indicate the start from April 12, 2020, the outbreak of the coronavirus disease (COVID-19) had been confirmed in around 209 countries or territories. The virus had infected 1,771,514 people overall the world, and the number of deaths had totaled 108,503. And also total cured or recovered cases are 402,110. The most severely affected countries include the U.S., Italy, and Spain, etc.



	Date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-22	Albania	0	0	0
2	2020-01-22	Algeria	0	0	0
3	2020-01-22	Andorra	0	0	0
4	2020-01-22	Angola	0	0	0
...	...	...	...	...	...
14980	2020-04-11	West Bank and Gaza	268	57	2
14981	2020-04-11	Western Sahara	4	0	0
14982	2020-04-11	Yemen	1	0	0
14983	2020-04-11	Zambia	40	28	2
14984	2020-04-11	Zimbabwe	14	0	3

14985 rows × 5 columns

Figure 2: Indian COVID-19 Dataset

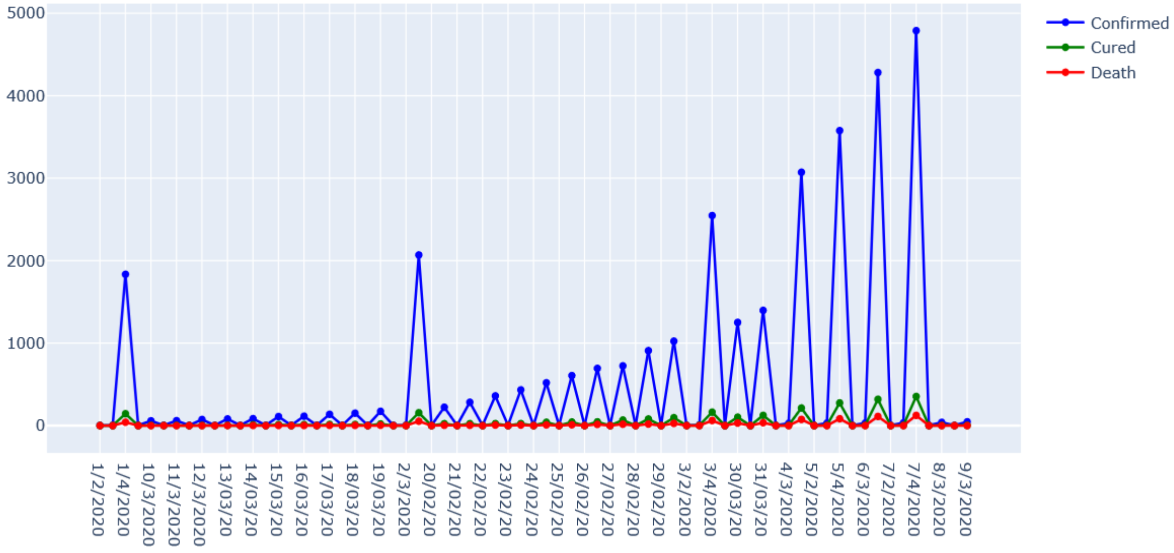


Figure 3: Visualizing of Indian COVID-19 Cases

Fig. 7 shows the current trends for the COVID-19 outbreak as displayed in Indian and the cases reported are visualized in the analytics dashboard to show the outbreak trend for confirmed, recovered, and death cases for all states. This aligns with the objectives of this paper show the outbreak progress over the period of time for each segment. It was found that the total number of confirmed cases for all states or regions is increasing steadily but on day 4/7/2020.

3. Current Outbreak Trends

The current trends for the COVID-19 outbreak as displayed in fig-9. The cases reported are visualized in the analytics dashboard to show the outbreak trend for confirmed, recovered, and death cases for all regions and countries. This aligns with our objectives to show the outbreak progress over the period of time for each segment. It was found that the total number of confirmed cases for all countries and regions is increasing rapidly, but at the end of March, the huge increments with 100,000 different cases every day approximately. The figs – show the confirmed cases, and recovered and also death cases exponentially especially the confirmed cases. Based on fig-9 the blue line indicates the confirmed cases, the green line indicates the recovered or cured cases, and also the red line indicates the death cases. One can observe the sharp rise (and falls, if any) very easily in this kind of visualization. It is useful for data with exponential relationships, or where one variable covers a large range of values. In this scenario, the case counts are increasing exponentially.

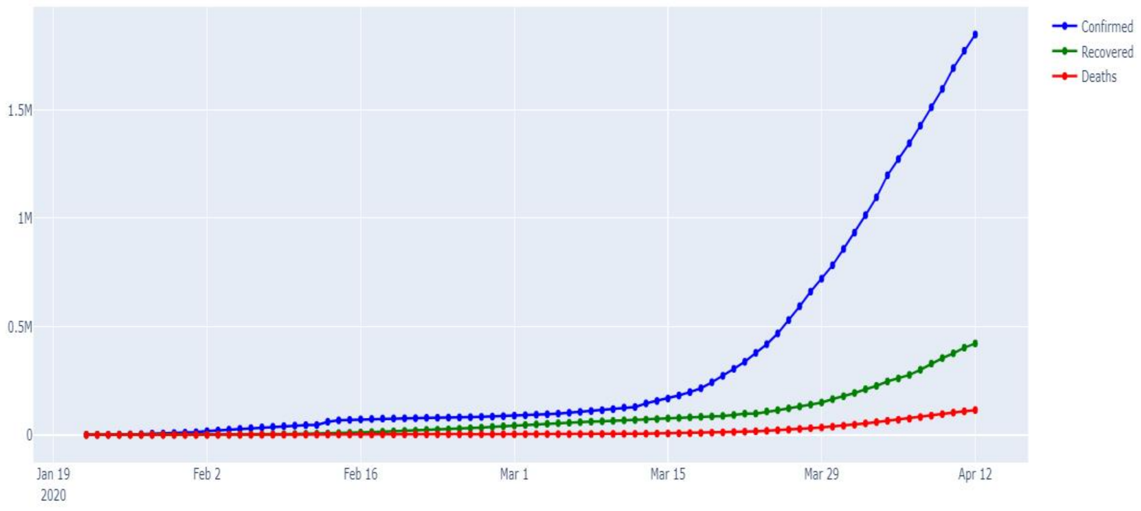


Figure 4: Visualizing of Worldwide COVID-19 Cases

The fig-5 is the data visualization of the till April 12/2020 COVID-19 cases spreading on the top ten countries as shown in the pie chart United States of America is 38.1 %. This is the highest affected country overall the world and the second top country is Spain 11.2 %, the third top country is 10.4 % and fourth in France 9.2% and also Germany is fifth-ranked among the top ten countries by 8.8 and also the last one is Belgium 2.0 % from the given top ten countries. This pie chart is indicates the top ten countries confirmed cases at the start from January 22/ 2020 till April 12/2020.

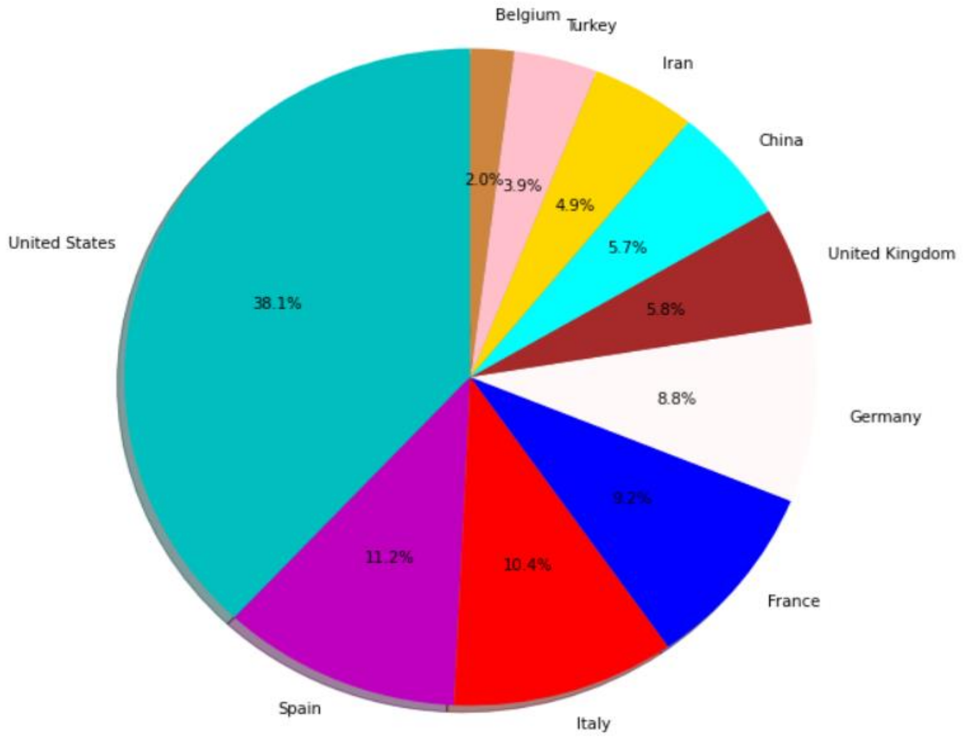


Figure 5: Top 10 Affected Countries

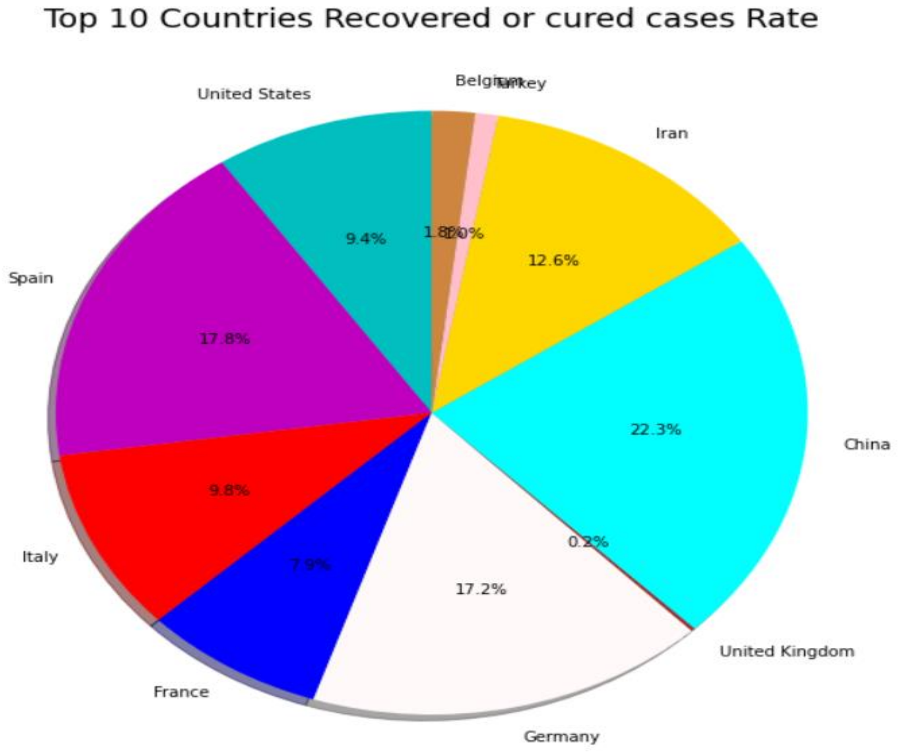


Figure 6: Recovered cases from the top 10 Countries

The fig-6 is the data visualization of the till April 12/2020 COVID-19 recovered cases overall top ten countries as shown in the pie chart China is 22.3% of people are recovered from COVID-19. And Spain 17.8 %, the third top country is 17.2 % and fourth is Iran 12.6% and also Italy is fifth-ranked among the top ten countries by 9.8% and also the last one in the United Kingdom 20.2 % from the given top ten countries. This pie chart is indicates the top ten countries recovered cases at the start from January 22/ 2020 till April 12/2020.

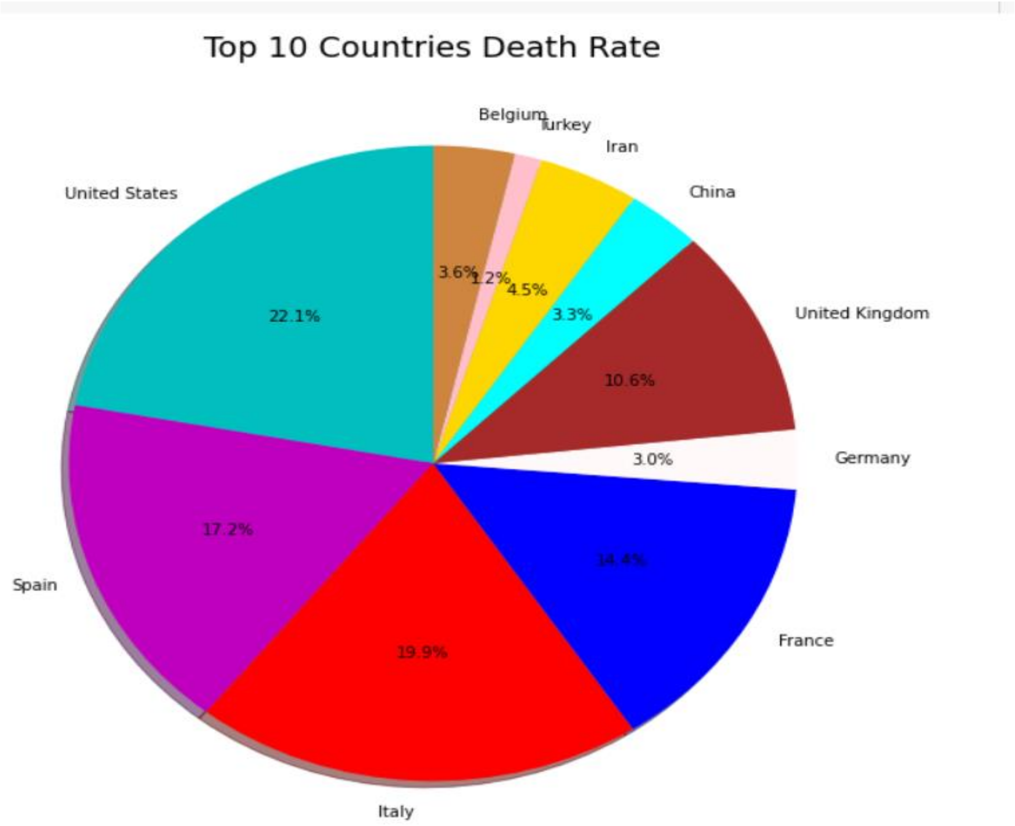


Figure 7: Death cases from the top 10 Countries

The fig-7 is the data visualization of the till April 12/2020 COVID-19 as we have seen from the pie chart starting from January 22/ 2020 till April 12 the death rate of the top ten countries. Based on the given graph analysis the highest death cases have happened in United States of America 21.1% overall the world and 19.9% death cases is happened in Italy, third 17.2% in Spain, 14.4% in France and 10.6% in the United Kingdom this highest than recovered cases in the UK. In China 3.3% and Turkey 1.2% of the top ten countries, this is fewer death rates when comparing other countries.

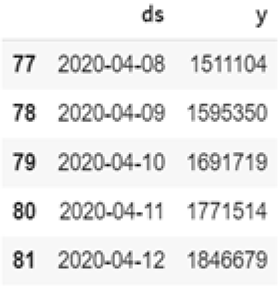
4. Predictive Modeling for future trend analysis

Just because the rise in number of cases is exponential, it does not imply that we can fit the data to an exponential curve and predict the number of cases in the coming days. Compartmental model techniques are normally used to model infectious diseases. Same could be used in the case of COVID-19 too. In fact, predictive analytics have been used in many and different sectors and industries such as manufacturing, education, market, and healthcare. As a matter of fact, predictive analytics is considered as an opportunity for the healthcare sector to be able to extract valuable information from data and predicting the future. Moreover, this opportunity can transform healthcare to not only predictive but to a preventive sector by the early detection of risks and the ability to make better decisions and saving more people's lives [7]. The predictive model does the analysis for identifying the patterns observed in historical and transactional data to predictive analytics comprises of several statistical and analytical techniques for developing strategies for the future possibilities of prediction. Therefore, Predictive analytics becomes vital when an essential quantity of highly sensitive data has to be handled. Based on the perceived events, future probabilities and measures are predicted. With the aid of available data mining techniques, predictive analytics predicts the events in the future and can make recommendations called prescriptive analytics [9]. Predictive analytics and data mining uses algorithms to discover knowledge and find the best solutions. Data mining is a process based on algorithms to analyze and extract useful information and automatically discover hidden patterns and relationships from data. In this phase I will predict the future impacts of the COVID-19 virus, the confirmed, recovered and death cases overall the world based on the current dataset. There some Data mining perdition and analysis algorithms are there. For this paper, I have proposed the time series prediction algorithm by using fbprophet python library to forecast the estimation of affected people, recovered and deaths for the coming 10 days, or the future assumption of the COVID-19 virus. As we know in this time COVID-19 virus is increasing day today as the report of World Health Organization (WHO), Worldometer and John Hopkins University, especially the confirmed is increasing the United States and some European countries. COVID-19 spreads globally in most of the countries and was defined pandemic by the WHO in March 2020 [10]. As of April 12, 2020, COVID-19 is affecting more than 186 countries and territories around the world with more than 1,846,679 confirmed cases, 421,722 recovered cases, and 114,091 deaths. The prediction models can help in health resources management and planning for prevention purposes. Google Search data is one of the information resources that contain useful information to predict and estimate epidemics. Data mining algorithms and techniques are well-known tools for predictive model development and data analysis.

4.1. Time Series Prediction and Analysis Algorithm

To forecast confirmed cases of COVID-19, we adopt simple time series forecasting approaches. I produce forecasts using models from the exponential smoothing family. This family has shown good forecast accuracy over several forecasting competitions and is especially suitable for short series. Exponential smoothing models can capture a variety of trend and seasonal forecasting patterns (such as additive or multiplicative) and combinations of those. I limit our attention to trended and non-seasonal models, given the patterns observed in fig-8.The objective of a predictive model is to estimate the value of an unknown variable. A time series has the time (t) as an independent variable (in any unit you can think of) and a target-dependent variable. The output of the model is the predicted value for y at time t and the minimum and maximum prediction confirmed cases, recovered cases, and death cases of the affected countries start from January 22/ 2020 till April 12/2020 to predict and analyze for COVID-19 virus for the coming 10 days. For this purpose, I have used fbprophet python library, and this is open-source software that is developed and released by a Facebook data scientist for the time series prediction and unseasonal situation. So, COVID-19 is not a seasonal issue because of this thing I have used this algorithm to implement this research. The forecasts (and 95% prediction intervals) produced at the end of 12/04/2020 and the mean estimate (point forecast) for the confirmed cases ten-days-ahead was 209 thousand with the 95% prediction intervals an absolute percentage error of 5%), with the forecasts being extremely positively biased. Still, the actual cases lie within the prediction intervals.

Table 4: confirmed case

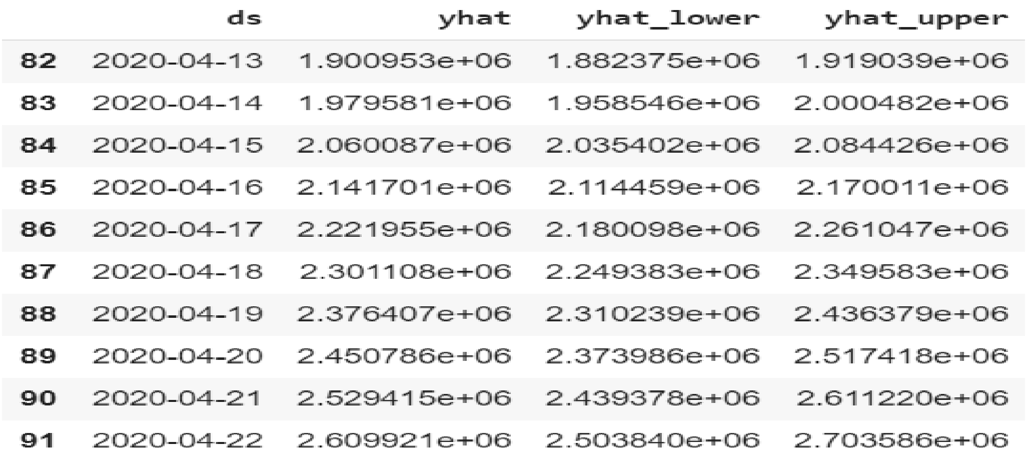


	ds	y
77	2020-04-08	1511104
78	2020-04-09	1595350
79	2020-04-10	1691719
80	2020-04-11	1771514
81	2020-04-12	1846679

As we have seen the table-4 this confirmed dataset starts from March 08/2020 to March 12/2020 and ds indicates DateTime stamp and y indicates the values in numeric form.

4.1.1. Forecasting of COVID-19 Confirmed cases Worldwide with Prophet base **model**

Table 5: forecasting of Confirmed cases



	ds	yhat	yhat_lower	yhat_upper
82	2020-04-13	1.900953e+06	1.882375e+06	1.919039e+06
83	2020-04-14	1.979581e+06	1.958546e+06	2.000482e+06
84	2020-04-15	2.060087e+06	2.035402e+06	2.084426e+06
85	2020-04-16	2.141701e+06	2.114459e+06	2.170011e+06
86	2020-04-17	2.221955e+06	2.180098e+06	2.261047e+06
87	2020-04-18	2.301108e+06	2.249383e+06	2.349583e+06
88	2020-04-19	2.376407e+06	2.310239e+06	2.436379e+06
89	2020-04-20	2.450786e+06	2.373986e+06	2.517418e+06
90	2020-04-21	2.529415e+06	2.439378e+06	2.611220e+06
91	2020-04-22	2.609921e+06	2.503840e+06	2.703586e+06

The above table shows the forecasting of COVID-19 virus worldwide starting from April 12/2020 till April 22/2020 yhat indicates the exact prediction, yhat_lower is the minimum prediction and yhat_upper is the maximum prediction for the coming 10 days. So based on the above table unto April 22/2020 the confirmed cases are increasing to 2,609,921. As we absorbed from the table day-by-day approximately 100,000 people are affected and it increases exponentially both the actual value and predicted values especially start from 24/03/2020 till 07/04/2020 the predicted values is increase a little bit than the actual values. These models help us evaluate or thing over it how much it difficult and spreading day-by-day the depth of the situation and plan for the worst. When it comes to saving lives, it is very important to consider worst-case scenarios and plan for them because the cost incurred to society during the worst-case scenario is significantly higher than the average case scenario.

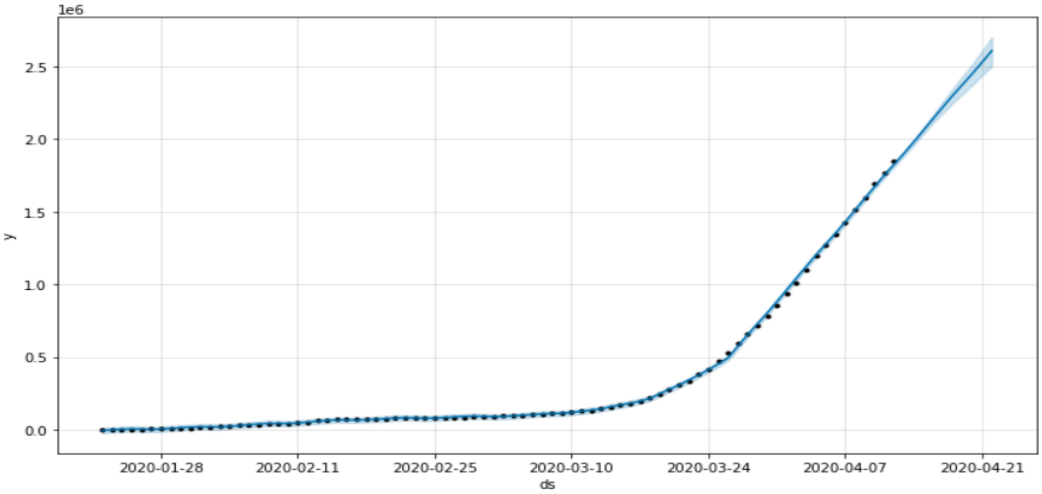


Figure 8: the predicted and original values

The fig-8 indicates the relationship of the original values that mentioned fig-4 the blue dotted line and the solid line is indicates the predicted values of the COVID-19 confirmed cases overall affected countries. Almost the predicted and the original values are going in the same way as we have seen in fig-8.

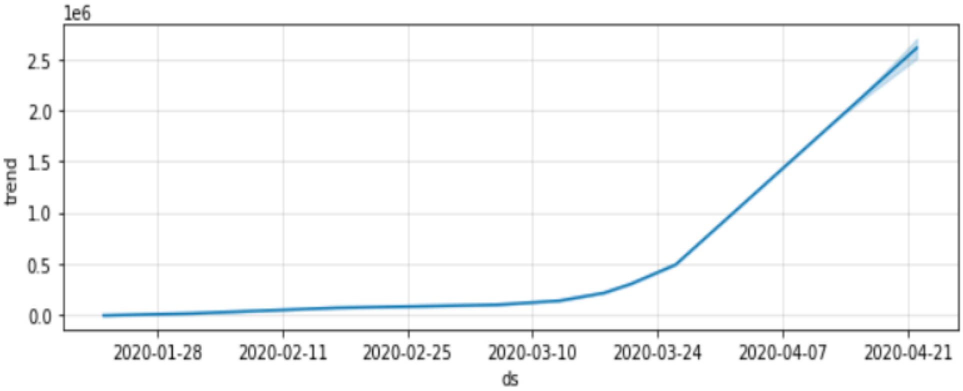


Figure 9: the predicted values of confirmed cases.

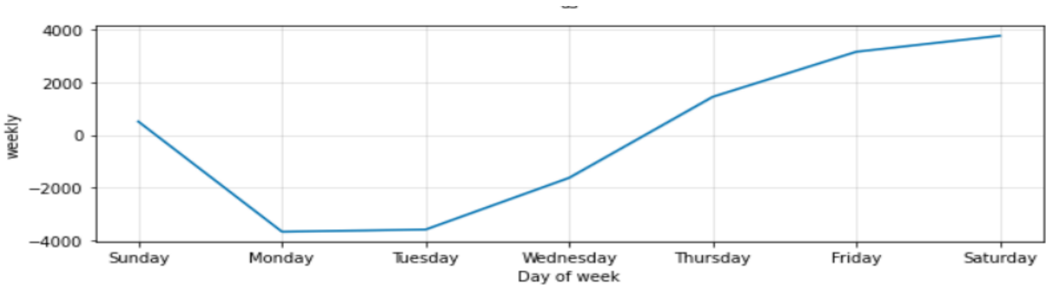
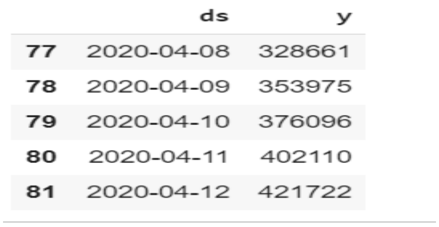


Figure 10: the weekly analysis of confirmed cases.

Fig-9 and 10 indicate the trend of confirmed cases and weekly analysis of COVID-19 cases till April 22/2020. From 28/01/2020 till 11/02/2020 the weekly analysis was decreasing and after Tuesday as we have seen in the fig-10 it increases again.

Table 6: Recovered cases.

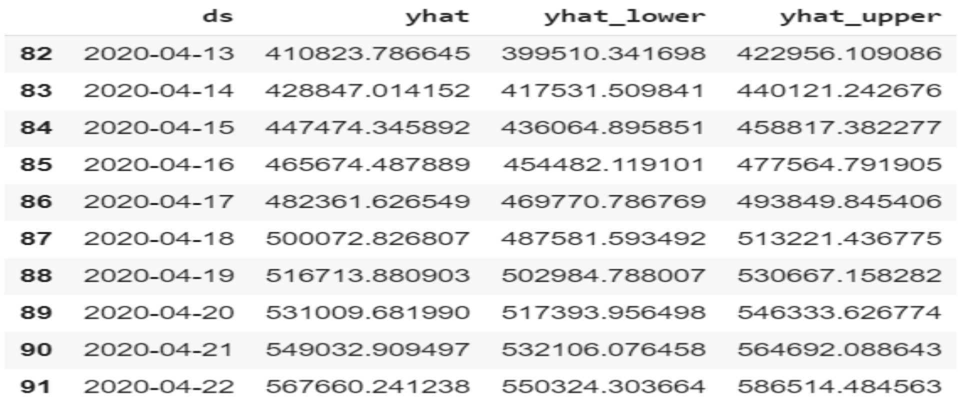


	ds	y
77	2020-04-08	328661
78	2020-04-09	353975
79	2020-04-10	376096
80	2020-04-11	402110
81	2020-04-12	421722

As we have seen the table-6 this recovered dataset starts from March 08/2020 to March 12/2020 and ds indicates DateTime stamp and y indicates the values in numeric form.

4.1.2. Forecasting of COVID-19 Recovered cases Worldwide with Prophet base model

Table 7: Prediction of recovered cases



	ds	yhat	yhat_lower	yhat_upper
82	2020-04-13	410823.786645	399510.341698	422956.109086
83	2020-04-14	428847.014152	417531.509841	440121.242676
84	2020-04-15	447474.345892	436064.895851	458817.382277
85	2020-04-16	465674.487889	454482.119101	477564.791905
86	2020-04-17	482361.626549	469770.786769	493849.845406
87	2020-04-18	500072.826807	487581.593492	513221.436775
88	2020-04-19	516713.880903	502984.788007	530667.158282
89	2020-04-20	531009.681990	517393.956498	546333.626774
90	2020-04-21	549032.909497	532106.076458	564692.088643
91	2020-04-22	567660.241238	550324.303664	586514.484563

The above table-7 shows the forecasting of COVID-19 virus worldwide starting from April 12/2020 till April 22/2020 yhat indicates the exact prediction, yhat_lower is the minimum prediction and yhat_upper is the maximum prediction for the coming 10 days. So based on the above table till April 22/2020 the recovered cases are increasing to 567,660.24

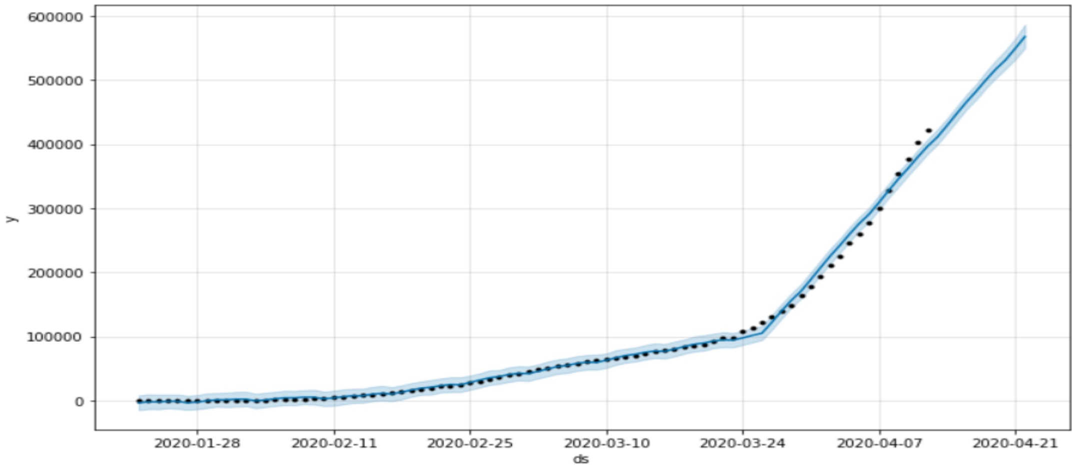


Figure 11: the relationship of the predicted and original values

The fig-11 indicates the relationship of the original values that mentioned the fig-4 the green dotted line and solid blue line indicates the predicted values of the COVID-19 recovered cases overall affected countries. Almost the predicted and the original values are going in the same way but after 07/04/2020 the original values increasing than the predicted values and also start from 24/03/2020 the predicted value is increased than the original values till 07/04/2020 as we have seen in fig-11.

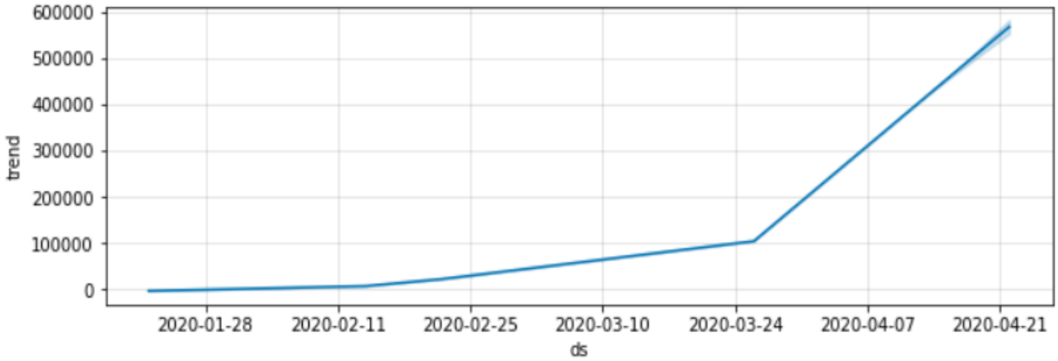


Figure 12: the predicted values of recovered cases

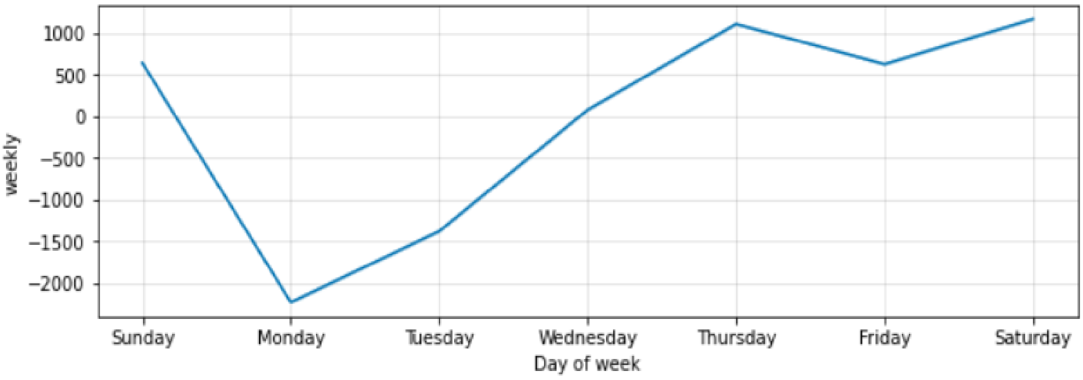
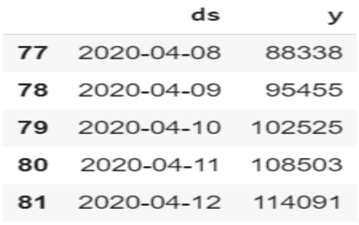


Figure 13: Weekly analysis of recovered cases

Fig-12 and 13 indicates the trend of recovered cases and weekly analysis of COVID-19 cases till April 22/2020. From 28/01/2020 till 11 /02/2020 the weekly analysis it was decreasing and after Monday as we have seen in the fig-10 it increases again till 24/03/2020 and also it is decreasing till 07/04/2020 and again after 07/04/2020 it also increasing.

Table 8: death cases

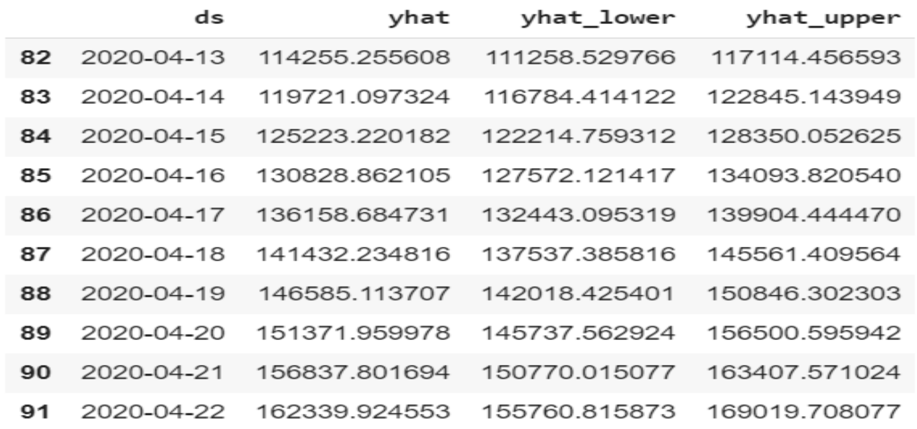


	ds	y
77	2020-04-08	88338
78	2020-04-09	95455
79	2020-04-10	102525
80	2020-04-11	108503
81	2020-04-12	114091

As we have seen the table-8 this death dataset start from March 08/2020 to March 12/2020 and ds indicates DateTime stamp and y indicates the values in numeric form.

4.1.3. Forecasting of COVID-19 Death cases Worldwide with Prophet base model

Table 9: the predicted vales death cases



	ds	yhat	yhat_lower	yhat_upper
82	2020-04-13	114255.255608	111258.529766	117114.456593
83	2020-04-14	119721.097324	116784.414122	122845.143949
84	2020-04-15	125223.220182	122214.759312	128350.052625
85	2020-04-16	130828.862105	127572.121417	134093.820540
86	2020-04-17	136158.684731	132443.095319	139904.444470
87	2020-04-18	141432.234816	137537.385816	145561.409564
88	2020-04-19	146585.113707	142018.425401	150846.302303
89	2020-04-20	151371.959978	145737.562924	156500.595942
90	2020-04-21	156837.801694	150770.015077	163407.571024
91	2020-04-22	162339.924553	155760.815873	169019.708077

The above table-9 shows the forecasting of COVID-19 virus worldwide starting from April 12/2020 till April 22/2020 yhat indicates the exact prediction, yhat_lower is the minimum prediction and yhat_upper is the maximum prediction of death cases for the coming 10 days. So based on the above table till April 22/2020 the recovered cases are increasing to 162,339.92.

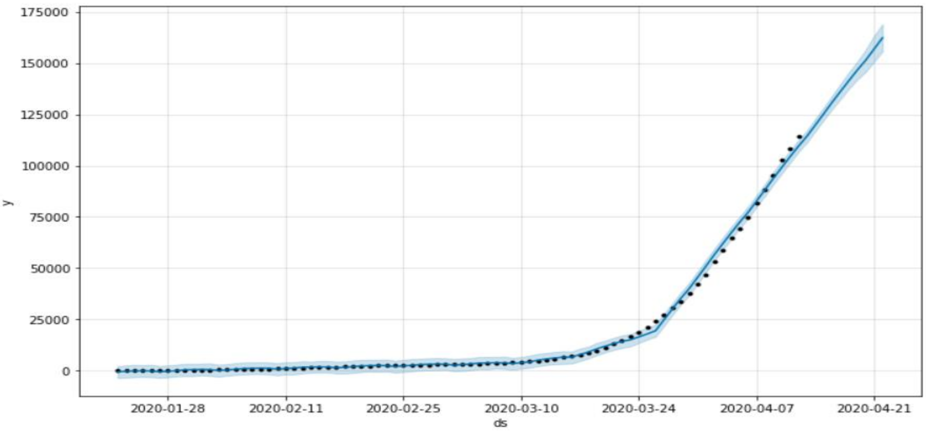


Figure 14: the relationship of the predicted and original values of death cases

The fig-14 indicates the relationship of the original values that mentioned the fig-4 the red dotted line and the solid blue line indicates the predicted values of the COVID-19 death cases overall affected countries. Almost the predicted and the original values are going in the same way but after 24/03/2020 the predicted values increasing than the original values till 07/04/2020 and also start from 07/04/2020 the original value is increasing than the predicted values as we have seen in fig-14.

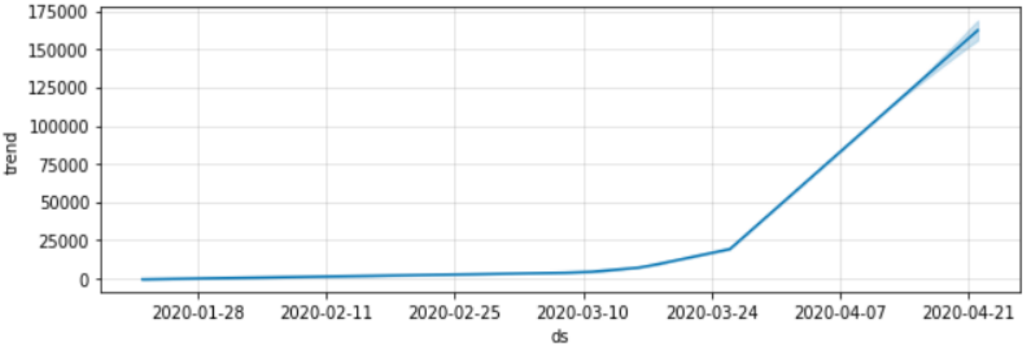


Figure 15: Trend of death cases

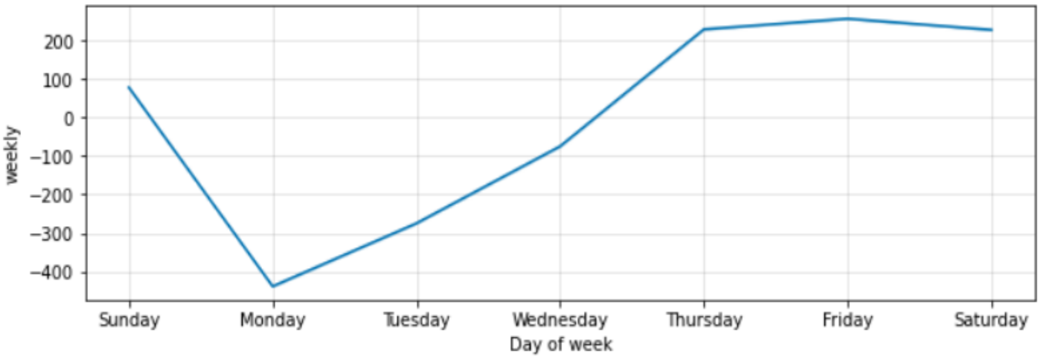
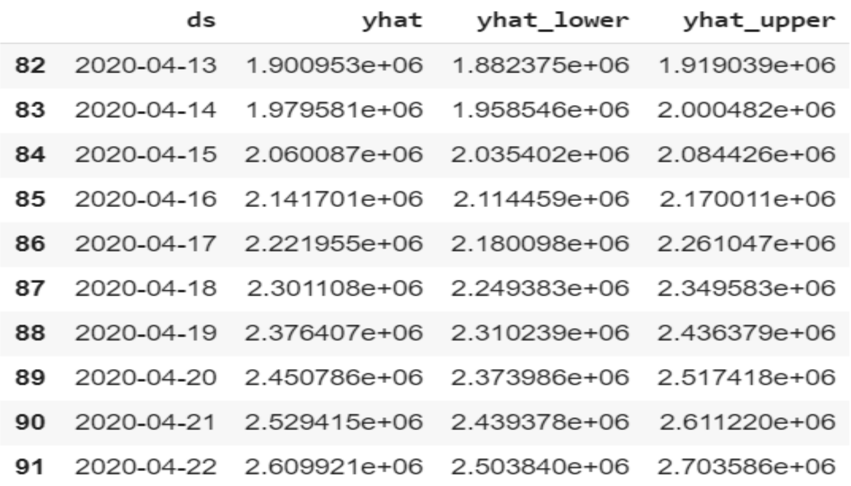


Figure 16: Weekly analysis of death cases

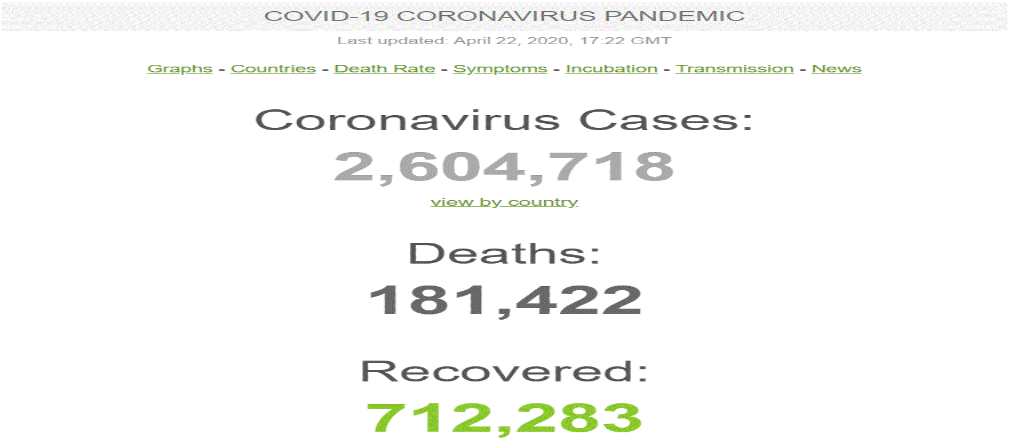
Fig-15 and 16 indicate the trend of death cases and weekly analysis of COVID-19 cases till April 22/2020. From 28/01/2020 till 11/02/2020 the weekly analysis was decreasing and after Monday as we have seen in the fig-16 it increases again till 24/03/2020 and after it is not increasing or decreasing but after 22/04/2020 it looks like decreasing.

5. Results and Discussion

The coronavirus disease has terrifically affected lives of people around the globe. Many people have lost their loved ones with the number of deaths worldwide currently goes beyond 100,000 keeps increasing exponentially. While Different technologies have penetrated into our daily lives with many successes, they have also contributed to helping humans in the extremely tough fight against COVID-19. This paper has predicted a survey of COVID-91 spreading so far in the literature relevant to the COVID-19 crisis's responses and control strategies. This paper basically analysis the COVID-19 outbreak prediction and analysis based on the confirmed, recovered and death cases on the given dataset that I took from WHO, Worldometer also GitHub.com, and DingXiangYuan and predict for the last 10 days by using time series data mining techniques. Based on the last 10 days predicted values and current status of COVID-19 outbreak status to show the result.



	ds	yhat	yhat_lower	yhat_upper
82	2020-04-13	1.900953e+06	1.882375e+06	1.919039e+06
83	2020-04-14	1.979581e+06	1.958546e+06	2.000482e+06
84	2020-04-15	2.060087e+06	2.035402e+06	2.084426e+06
85	2020-04-16	2.141701e+06	2.114459e+06	2.170011e+06
86	2020-04-17	2.221955e+06	2.180098e+06	2.261047e+06
87	2020-04-18	2.301108e+06	2.249383e+06	2.349583e+06
88	2020-04-19	2.376407e+06	2.310239e+06	2.436379e+06
89	2020-04-20	2.450786e+06	2.373986e+06	2.517418e+06
90	2020-04-21	2.529415e+06	2.439378e+06	2.611220e+06
91	2020-04-22	2.609921e+06	2.503840e+06	2.703586e+06



COVID-19 CORONAVIRUS PANDEMIC

Last updated: April 22, 2020, 17:22 GMT

Graphs - Countries - Death Rate - Symptoms - Incubation - Transmission - News

Coronavirus Cases:

2,604,718

view by country

Deaths:

181,422

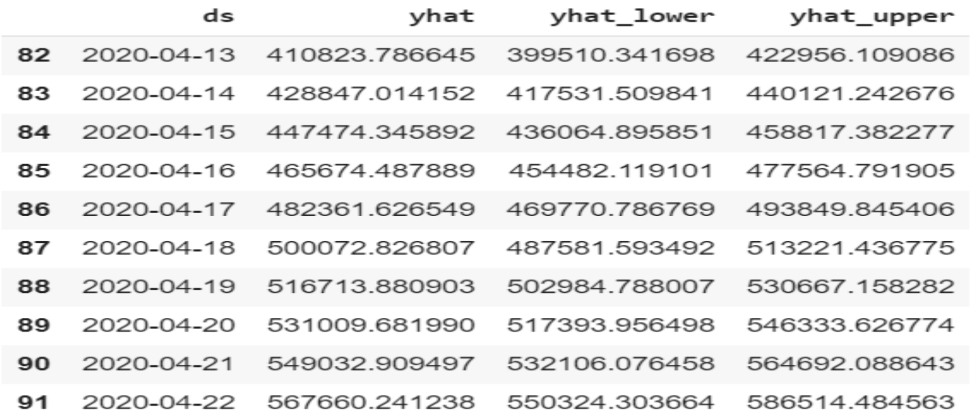
Recovered:

712,283

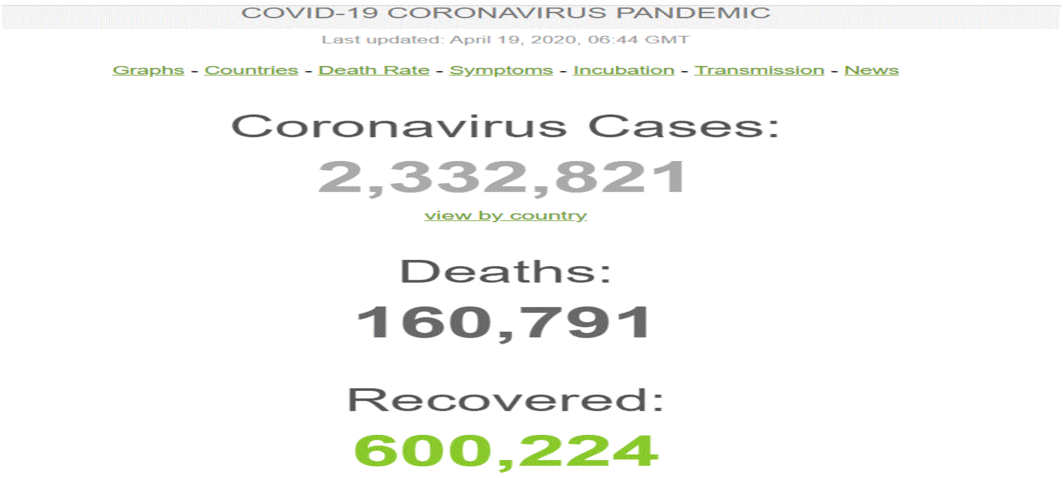
Figure 17: COVID-19 status

When we have seen the confirmed case predicted values start from April 12/2020 to April 22/2020 the predicted value is mentioned as follows below tables. The predicted values of the Confirmed case are achieved 99% on that exact date means April 22/2020 as we have seen from the predicted table on the day of April 22/2020 is the lower prediction is 2,503,840.06 normal prediction is 2,609,921.06 and when we see the heights prediction 2,703,5586.06. Worldometr report on April 22/04/2020 COVID-19 status is 2,604,718 confirmed cases as we have seen on the fig-17 as overall the world.

The predicted values of the recovered case are achieved 99% before predicted date as we have seen from the predicted table on the day of April 22/2020 is the lower prediction is 550,324.30, the normal prediction is 567,660.24 and when we see the heights prediction 586,515.48. So The Worldometr report on April 19/2020 status is 600,224 recovered cases as we have seen on the fig-18 as overall the world.



	ds	yhat	yhat_lower	yhat_upper
82	2020-04-13	410823.786645	399510.341698	422956.109086
83	2020-04-14	428847.014152	417531.509841	440121.242676
84	2020-04-15	447474.345892	436064.895851	458817.382277
85	2020-04-16	465674.487889	454482.119101	477564.791905
86	2020-04-17	482361.626549	469770.786769	493849.845406
87	2020-04-18	500072.826807	487581.593492	513221.436775
88	2020-04-19	516713.880903	502984.788007	530667.158282
89	2020-04-20	531009.681990	517393.956498	546333.626774
90	2020-04-21	549032.909497	532106.076458	564692.088643
91	2020-04-22	567660.241238	550324.303664	586514.484563



COVID-19 CORONAVIRUS PANDEMIC

Last updated: April 19, 2020, 06:44 GMT

Graphs - Countries - Death Rate - Symptoms - Incubation - Transmission - News

Coronavirus Cases:

2,332,821

Deaths:

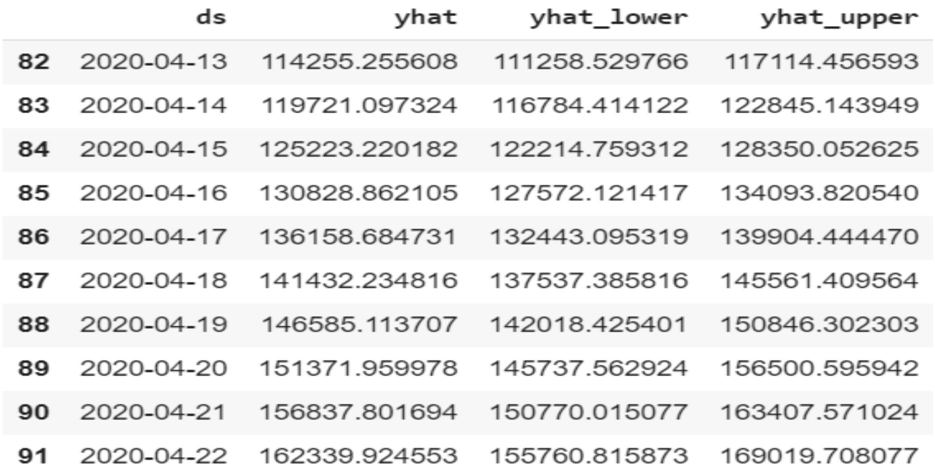
160,791

Recovered:

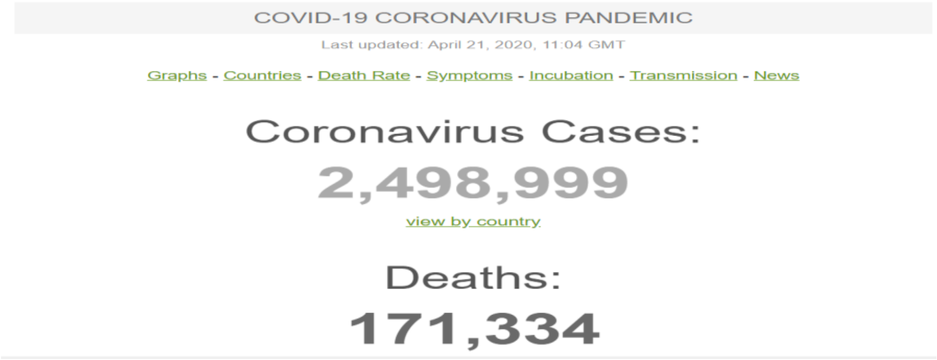
600,224

Figure 18: COVID-19 result

The predicted values of the death case are achieved 99% before predicted date as we have seen from the predicted table on the day of April 22/2020 is the lowest prediction is 155,760.82, the normal prediction is 162,339.92 and when we see the heights prediction 169,019.71. So The Worldometr report on April 21/2020 status is 171,334 recovered cases as we have seen on the fig-19 as overall the world.



	ds	yhat	yhat_lower	yhat_upper
82	2020-04-13	114255.255608	111258.529766	117114.456593
83	2020-04-14	119721.097324	116784.414122	122845.143949
84	2020-04-15	125223.220182	122214.759312	128350.052625
85	2020-04-16	130828.862105	127572.121417	134093.820540
86	2020-04-17	136158.684731	132443.095319	139904.444470
87	2020-04-18	141432.234816	137537.385816	145561.409564
88	2020-04-19	146585.113707	142018.425401	150846.302303
89	2020-04-20	151371.959978	145737.562924	156500.595942
90	2020-04-21	156837.801694	150770.015077	163407.571024
91	2020-04-22	162339.924553	155760.815873	169019.708077



COVID-19 CORONAVIRUS PANDEMIC

Last updated: April 21, 2020, 11:04 GMT

Graphs - Countries - Death Rate - Symptoms - Incubation - Transmission - News

Coronavirus Cases:

2,498,999

view by country

Deaths:

171,334

Figure 19: Result of Death rate.

6. Recommendations for the Future

Currently, the whole world is looking forward to data-driven insights to combat coronavirus. Data analytics has already started to show fruitful results by predicting the impact of coronavirus. To devise an appropriate predictive analytics model for analyzing the outbreak, mining medical data isn’t sufficient, but you also need to get faster and smarter at it. By leveraging Quantizing’s advanced data analytics solutions that leverage data mining and text analytics you will be able to convert bigger datasets into actionable insights in a much smaller span of time. We must take note of this fact that the healthcare industry doesn’t revolve around just physicians and patients but also third parties like insurance companies. Secured and organized data is a good opportunity for organizations to save money. All they need is to represent their facilities in an accurate way by leveraging advanced data warehousing and data mining solutions The data mining algorithm is very sensitive are of the COVID-19 and other related sectors to predict and forecast such like dangers and difficult virus for upcoming times and announce for the people to take pre-controlling mechanism for their life. In future, researchers should focus on data collection and sentiment analysis by using machine learning algorithm will work a lot better with bigger local datasets. Also, the local datasets are more accurate than the datasets available on the internet. So, researchers should find a way to collect big data in the future. The data will significantly improve the findings of this research. The area of text classification has attracted a lot of interest from both the machine learning research community and the industry, one the popular application of text classification is sentiment analysis, whose objective is to guess the positive or negative attitude of a user towards a topic give a sentence, I will give an overview of how to apply machine learning techniques to text classification and sentiment analysis. For future work I will recommend sentiment analysis of the people's sentiment means positive and negative filling about COVID-19 virus by collecting data form different social media and related data sources by using machine learning approach to classifying the negative and positive impacts of COVID-19 virus. As very interesting business application of text classification is sentiment analysis is a method to automatically understand the perception of people towards the COVID-19 based on their comment. The inputs text is classified into positive, negative and in some situations, neural. As we know Social media offers a powerful outlet for peoples thoughts and feelings – it is an enormous ever-growing source of texts ranging from everyday observations to involved discussions. This thesis contributes to the field of sentiment analysis, which aims to extract emotions and opinions from text to analysis and classifying accordingly by the help of classification algorithm and sentiment analysis.

7. Conclusion

In this work, I have shown the prediction of covid19 by using a time series data mining technique based on the current dataset on the proposed combination of three major pillars to analyze the outbreak of the COVID-19 virus. The coronavirus disease has terrifically affected the lives of people around the globe. Many people have lost their loved ones with the number of deaths worldwide currently goes beyond approximately 100,000 in 24 hours increasing. While Data mining techniques and related technologies have penetrated into our daily lives with many successes, they have also contributed to helping humans in the extremely tough fight against COVID-19. This paper has presented a predicted and analysis of confirmed, recovered, and death cases of COVID-19 and forecasting based on the number of cases time series based on the current data. Although various studies have been published, we observe that there are still relatively limited applications and contributions of Data mining in this battle. This is partly due to the limited availability of data about COVID-19 whilst Data mining methods normally require large amounts of data for computational models to learn and acquire knowledge. However, we expect that the number of Data mining studies and research areas related to COVID-19 and other things increase significantly and play great roll for people especially these kinds of worst time. This paper mainly predicted the COID-19 outbreak for the last 10 days and analysis graphically by using the data mining time series technique for both confirmed, recovered and death cases. When we see the perdition it achieved 99%, for the three cases means confirmed, recovered and death cases. The Constraints/limitation of this research is it was difficult to get related works because as we know COVID-19 virus is new virus. So, I could not get organized articles and related sources.

Acknowledgments

Firstly, I want to express my gratitude to Almighty God who gave me power and patience in every endeavor of my life. Next, I express my privilege to express my sincerest regards to my guide, Dr. Deepak Kumar - Assistant professor CSE National Institute of Meghalaya, for his valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration. I pay my respects and love to my parents and all other family members and friends for their love and encouragement throughout my career. Last but not the least we express my thanks to my friends for their cooperation and support. And Finally I would like to thank you Mr. Abhishek Chand for his support and helping by different things from the starting end of my research work.

Reference

[1]. Saima Bano, Muhammad Naeem Ahmed Khan,“ A Framework to Improve Diabetes Prediction using k-NN and SVM”, International Journal of Computer Science and Information Security (IJCSIS),Vol. 14, No. 11,pp.3-10, November 2016.

[2]. Behrouz Pirouz, Sina Shaffiee Haghshenas, Patrizia Piro2, “Investigating a Serious Challenge in the Sustainable Development Process: Analysis of Confirmed cases of COVID-19 (New Type of Coronavirus) Through a Binary Classification Using Artificial Intelligence and Regression Analysis”, 5 March 2020.

[3]. Saima Bano, Muhammad Naeem Ahmed Khan,“ A Framework to Improve Diabetes Prediction using k-NN and SVM”, International Journal of Computer Science and Information Security (IJCSIS),Vol. 14, No. 11, November 2016.

[4]. Corona Tracker Community Research Group, “World-wide COVID-19 Outbreak Data Analysis and Prediction” ,No.4, 19 March 2020.

[5]. Desmond Bala Bisandu1, Dorcas Dachollom Datiri2 , Eva Onokpasa3 , Godwin Thomas4 , Musa Maaji Haruna5 , Aminu Aliyu6 , Jerry Zachariah Yakubu7, “Diabetes Prediction Using Data Mining Techniques”, International Journal of Research and Innovation in Applied Science (IJRIAS) | Volume IV, Issue VI, June 2019|ISSN 2454-6194.

[6]. B. Kavitha Rani1 and A.Govardhan2, “Rainfall Prediction Using Data Mining Techniques - A Survey”, DOI: 10.5121/csit.2013.3903.

[7]. Fatimetou Zahra Mohamed Mahmoud, “The application of predictive analytics in healthcare sector”, International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 8, August 2017.

[8]. Corona Tracker Community Research Group, “World-wide COVID-19 Outbreak Data Analysis and Prediction” ,No.4, 19, pp.8-18, March 2020.

[9]. S. Poornima* and M. Pushpalatha, “ A survey of predictive analytics using big data with data mining”, Int J of Bioinformatics Research and application, Vol. 14, No.3, 2018.

[10]. Seyed Mohammad Ayyoubzadeh, Seyed Mehdi Ayyoubzadeh, Hoda Zahedi,Mahnaz Ahmadi, Sharareh R. Niakan Kalhori, “Predicting COVID-19 Incidence Using Google Trend and Data Mining Techniques: A case study of Iran (Preprint)”, March 21, 2020.