2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377198

Predicting the Prevalence Rate of COVID-19 Falsity on Temperature

Ananya Rakhra
Amity School of Engineering and Technology
Amity University
Uttar Pradesh, India
ananyarakhra@gmail.com

Raghav Gupta
Amity School of Engineering and Technology
Amity University
Uttar Pradesh
raghavgupta1@hotmail.com

Ishika Jain
Amity School of Engineering and Technology
Amity University
Uttar Pradesh, India
jain.ishika65@gmail.com

Madhulika Bhatia
Amity School of Engineering and Technolgy
Amity University
Uttar Pradesh, India
mbhadauria@amity.edu

***Abstract*—COVID-19 originally known as Corona virus has been declared as pandemic by the World Health Organization on 11th March 2020. This infectious disease discovered from Wuhan, China in December 2019 and has affected millions of people around the world. Every country around the world is undergoing global economic crises and therefore, it' s the need of an hour to predict the prevalence and incidence of this disease throughout the world. This will help the medical practitioners and government agencies in India to make key decisions and appropriate measures to demystify the disease and prevent the country from global economic recession. This paper aims to analyze the number of cases in India by utilizing the machine learning techniques and exploratory data analysis to observe the growth patterns and map the increase in the frequency of those infected. The source of data was authentic COVID-19website which was showing confirmed diseased cases of Delhi, Uttar Pradesh and India as a whole. The count of confirmed cases taken from 14th March 2020 to 3rd September 2020 put together will help to know how effective the current efforts have been and also help to realize the need of working further to combat this virus. This research focuses on predicting the possible number of confirmed cases using techniques of data mining, data analysis with particularly regression, clustering and predictive analysis. The primary focus is to predict the number of cases in the coming month and finding out that whether there is relation between temperature with number of confirmed cases or not.**

Keywords— Pandemic, Covid-19, Analysis, Prediction, Data mining, Regression, Clustering

I. Introduction

Covid-19 is an infectious disease caused by severe acute respiratory syndrome corona virus 2 (SARS-COV-2) [1]. The recent pandemic of covid-19 has come up as a global challenge [2]. It's consequences on the world's health care, economy, population are massive. Corona virus is a chain of viruses causing illness ranging from common cold to more severe diseases.

Data mining is the technique of predicting the possibility of a certain event or happening based on the current available data and resources [3]. The main use of this tool is to forecast or make predictions. It is basically organized data that helps to interpret and further elaborate on the given context. It can also be said that data mining is a technique of discovering patterns amidst large data sets involving methods that includes machine learning, statistics and database systems. Data mining techniques such as clustering, regression, sequential patterns and prediction, helping to get information based on knowledge. The data mining techniques are studied to learn more about predictive analysis based on different mediums.

COVID-19 has turned into a global challenge posing crisis like situations across the globe. For a developing country like India, the difficulties are further heightened where health care, economy and society are still under progress. The economy has been hit as an impact of unpreparedness to face a pandemic situation and alongside, healthcare facilities available did not match the requirement. Therefore, by predicting the expected number of cases in the near future based upon the patterns observed in the number of cases observed since it has entered our country will serve as a wakeup call regarding the bigger challenge that may come our way and give the time to ensure that we are ready to bring the situation under control.

Technique of data mining and prediction [4] helped the path to find that the number of cases in India may increase. Analysis also has been performed that is there any observable relationship exist between the temperature and the number of cases observed on any particular day. The technique of linear regression is applied to test relation in tools such as Rapid Miner and Weka.

Linear Regression is an approach utilized for finding estimate relationships [5] between a dependent variable and one or more independent variables. It predicts output with continuous values. In this project. The use of regression technique, to check whether there is a relation between temperature and increase or decrease of infected COVID-19 cases. It has been observed that temperature does not have an impact on the number of cases observed on any given day.

Data analysis is a technique of organizing, transforming and modeling data [6] with the aim of attaining information which is useful, yielding a conclusion and helping in the

process of decision making. The work has been collected and data is organized it into tables and charts to be able to yield meaningful results. Data analysis is performed is performed using different tools.

The contribution in this work intend to make through this project is very beneficial to the society especially in times of such unpredictable pandemic. Work shown the observation in the growth pattern of cases in our country over a period of 6 months (March to September) and on the basis of these observations prediction is done in the possible no. of cases up till February 21. This will benefit by allowing the society to be better prepared for the upcoming challenges which individuals may have to face, as during the initial outbreak of the pandemic, lack of preparedness caused hassle and widespread infection amongst the population. The contributions are listed as under:

1. Map the number of cases

2. Perform data analysis for better understanding and organization

3. Predict the cases in the following month

4. Use of data mining (regression and clustering) to find out any relationship between temperature and no. of cases considering data with gap of 10 days (from 14th March to 8th May).

The objective is to analyze the number of cases in our country i.e. India with the help of data analysis and data mining technique. Utilizing the technique of data analysis we will also observe its growth patterns and map the increase in the count of those infected. Using data mining and predictive analysis, we will predict the possible number of confirmed cases in the near future. This will helped to be better prepared for the upcoming challenges that we may have to face even under the worst case scenario. After observing the data related to confirmed cases in this country, the relation between COVID cases and temperature is analyzed.

II. Related Work

There is lot of work that has been studied in previous researches and it was found that the proposed work is different from traditional approaches.

The state of art techniques comprised of various topics and fields of interest. One of them was about the outbreak in Wuhan and its spread that followed [7]. Another one observed the spread through the United States of America [8]. We also came across a research paper that compared the case confirmation, recovery and death rate in China, Italy and France and observed severe similarity in the recovery trends [9]. A researcher also attempted to predict the number of cases in Italy assuming the pattern to be similar to China's outbreak trend. There have been articles about the prediction of epidemic peak in India [10]. Another researcher drew a comparison by predicting the number of cases in India with and without the lockdown [11]. One of the papers used time series forecasting, predicting the epidemic peak in April 2020, which is now proved inaccurate [12]. One paper was based on a regression model utilizing linear regression technique for prediction. Attempts have also been made by many researchers to establish relation between temperature, humidity and the number of cases in India as well as in the US [13].

This research holds similarity with the work mentioned above with the major improvement in studying the data mining techniques and using regression analysis for identifying the direct relationship with temperature as previously there were rumors that the increase in temperature is directly proportional to the reduction in the cases. The techniques of machine learning emphasized in this research will help to find the increased cases of the pandemic in the near future. In depth analysis has been taken forward to identify a relation between temperature and number of cases in India and the results are visualized through graph using data mining tools.

III. Methodology

The proposed work shows the use of data mining in for the prediction purpose. This research tries to predict, if there is an increase or decrease in no. of cases, in the coming month or not.

With the help of data analysis, the data is organized, transformed and modeled with the aim of attaining information which is useful, yielding a conclusion and helping in the process of decision making. The raw data has been processed to extract information. It makes easier for the user to manipulate and process data which infer results and decisions based on it. The research is performed on the basis of Data analysis on two attributes -qualitative and quantitative (experiments and surveys). Apart from this, the techniques included are- text analysis, statistical analysis, diagnostic analysis, predictive analysis and prescription analysis. This method has been used for collecting and organizing the data, related to COVID-19 cases, in India. The data has been presented in the form of tables and visualization of graphs are illustrated further in sections. The below Table 1 shows dates and number of cases reported. The data is shown in Table 1.

The complete methodology is shown in Fig 1:

The data of COVID-19 patients is taken from authentic source [14]. The data cleaning is performed to get the most relevant data. Data analysis is performed and results are shown in the form of varied visualization techniques. Data is analyzed for prediction using Weka software [15] by applying three techniques Regression and Clustering.

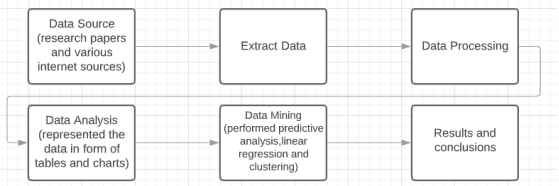

Figure 1: Methodology

IV. Results and Discussion

The disease has been spread globally and was declared as pandemic in March. In March 2020, India had the graph of increasing cases with high rate, Due to the analysis, the government had to take immediate actions of announcing the lockdown strategy initially for about 15-20 days [16]. Since the cases were increasing day by day, the lockdown had to get extended for more days which lead to months. In this research, the method of data analysis has been extensively applied, identifying the necessary benefits for predicting the uplifting of lockdown and other isolation strategies to prevent the spread of this pandemic as it is prone to human to human virus. Data is collected regarding number of confirmed cases in country and same is applied in the capital of India, where the population is the highest among other states. The data is cleaned and extracted to show number of confirmed cases in daily as well as its cumulative and daily figures of Delhi as well as shown in Fig. 2 and Fig. 3 respectively.

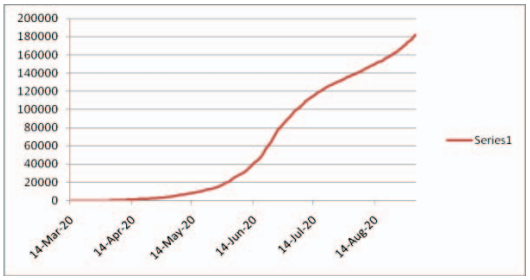

Figure 2: Confirmed cases in Delhi

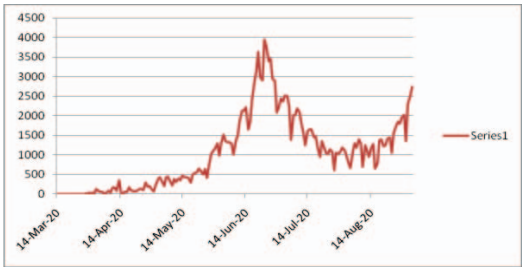

Figure 3: Daily cases confirmed in Delhi

The data extracted to observe the number of confirmed cases in Uttar Pradesh has been presented in graphs as below.

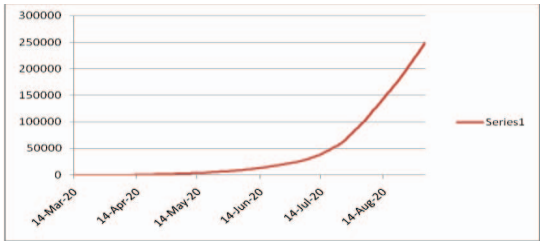

Figure 4: Confirmed cases in Uttar Pradesh

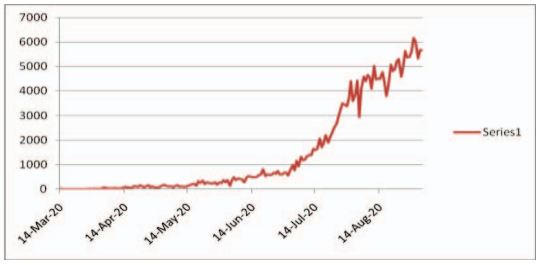

Figure 5: Daily cases confirmed in Uttar Pradesh

The analysis has been performed on data for two states of India which are, Delhi and Uttar Pradesh. The increased number of confirmed cases is studies in the whole country and illustrated using visualization tools. The prediction in the no. of cases in the coming month is achieved, with the help of data mining and predictive analysis techniques. Data for cases observed in India, cumulative and daily Fig. 5and 6 below:

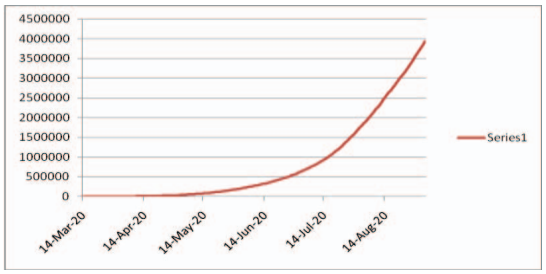

Figure 6: India cumulative cases

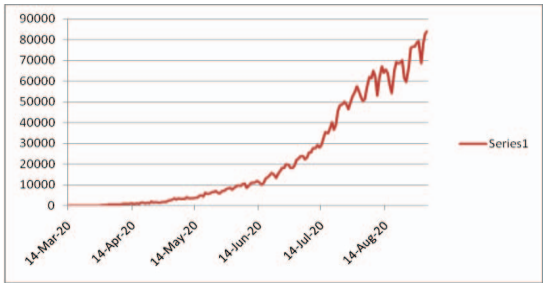

Figure 7: Daily cases of India

Estimates have been about the expected number of cases in India up till 24thFebruary□21. The values and forecasted values are analyzed along with lower and upper confidence board. The values are as presented below followed by the number of cases expected every day for the infected individuals. Prediction is applied on the data

collected in Tabular form to get below results predicting the exponential increase in number of Cases

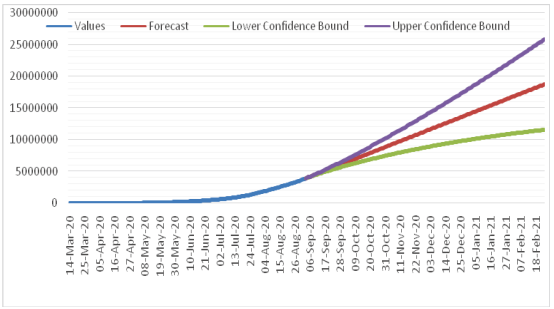

Figure 8: Values and forecasted values

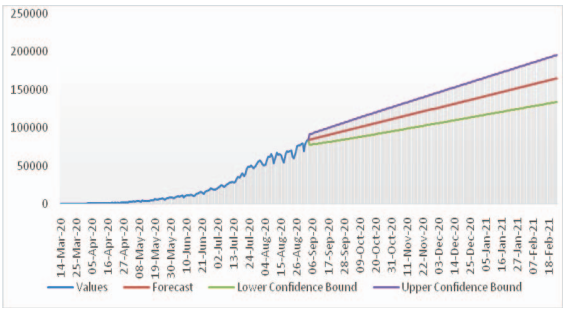

Figure 9: Daily cases values and forecast

The data is observed for a relation between the no. of cases and temperature in Delhi was. The data is imported in Weka tool [17]. To find out the relationship Regression is applied and clusters are generated. The regression model is generated around after the prediction, to find a relation between no. of COVID-19 cases in Delhi and the temperature. For this purpose the case data over the months of April and May have been utilized. The regression analysis is performed and the results are shown in Table 2.

Table 2: Regression Results

Correlation coefficient	Root mean squared error (RMSE)	Mean absolute error (MAE)
-0.9929	6.1319	4.675

There are 6 clusters and according to the table, from data six rows only. The clusters are used to find similarities (if any) between 6 instances (from 14th march to 8th may with a gap of 10 days). But there was no such relation or similarity found, as all the cluster instances were far apart and small in size, due to which there wasn't any clear image.

The exploratory data is analyzed and the results are shown in Fig 8.

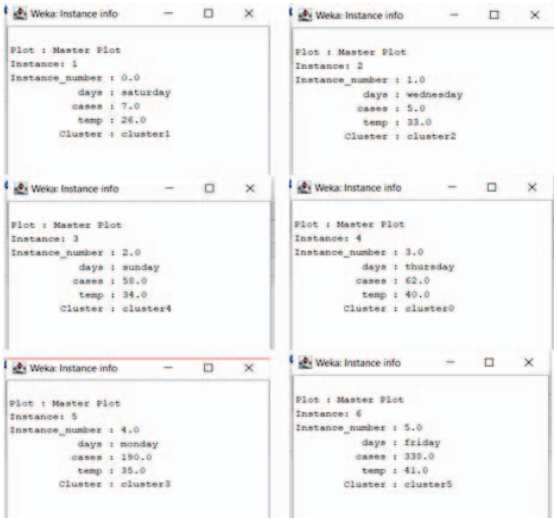

Figure 10: Cluster Instances

The clusters are each at different corners or places, so it is observed that no cluster is related to each other, the summary of clusters shown in Table 3

Table 3: Clusters summary

Instance no.	Clusters	Cases	Temperature
0.0	cluster1	7.0	26.0
1.0	cluster2	5.0	33.0
2.0	cluster4	58.0	34.0
3.0	cluster0	62.0	40.0
4.0	cluster3	190.0	35.0
5.0	cluster5	338.0	41.0

Cluster 3 shows that temperature decreased and the cases get increased as shown in Fig. 9.

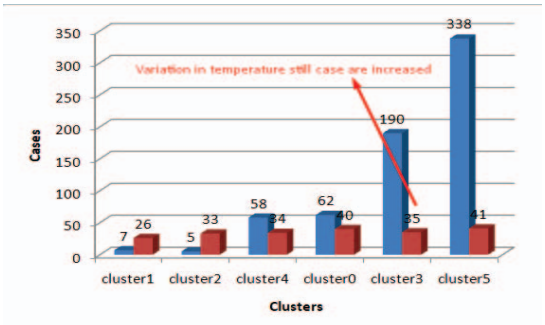

Figure 11: Variation in cases with temperature

V. Conclusion

The covid-19 virus has been declared as a pandemic by the WHO. The only path to overcome the same is when an effective vaccine tackles it. Through this research we forecasted the adversities that may be faced by us in the near future and allow us to better prepare for the same. . The research work performed has stated the prediction regarding prevalence of this infection in coming month, it also establishes falsity of the hypothesis that- there is any relation between temperature and the virus's positivity rate. The proven independency of positivity rate on temperature has helped clear of ambiguity and false hope for this pandemic to subside with the warmer seasons and therefore urges for preparedness towards it.

References

[1] Sun, Pengfei, Xiaosheng Lu, Chao Xu, Wenjuan Sun, and Bo Pan. "Understanding of COVID-19 based on current evidence." *Journal of medical virology* 92, no. 6 (2020): 548-551.

[2] World Health Organization. "Coronavirus disease 2019 (COVID-19): situation report, 72." (2020).

[3] Bhatia, Surbhi, Poonam Chaudhary, and NilanjanDey. "Introduction to Opinion Mining." In *Opinion Mining in Information Retrieval*, pp. 1-22. Springer, Singapore, 2020.

[4] Gulati, Hina. "Predictive analytics using data mining technique." In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 713-716. IEEE, 2015.

[5] Seber, George AF, and Alan J. Lee. *Linear regression analysis*. Vol. 329. John Wiley &Sons, 2012.

[6] Peng, Liangrong, Wuyue Yang, Dongyan Zhang, ChangjingZhuge, and Liu Hong. "Epidemic analysis of COVID-19 in China by dynamical modeling." *arXiv preprint arXiv:2002.06563* (2020).

[7] Bouighoulouden, Amine, and IlhamKissani. "CROP YIELD PREDICTION USING K-MEANS CLUSTERING." (2020).

[8] Gupta, Sonal, Gourav Singh Raghuwanshi, and Arnab Chanda. "Effect of weather on COVID-19 spread in the US: a prediction model for India in 2020." *Science of The Total Environment* (2020): 138860.

[9] Sajadi, Mohammad M., Parham Habibzadeh, Augustin Vintzileos, ShervinShokouhi, Fernando Miralles-Wilhelm, and Anthony Amoroso. "Temperature and latitude analysis to predict potential spread and seasonality for COVID-19." *Available at SSRN 3550308* (2020).

[10] Wagh, Chaitanya S., Parikshit N. Mahalle, and Sanjeev J. Wagh. "Epidemic peak for COVID-19 in India, 2020." (2020).

[11] Tiwari, Sunita, Sushil Kumar, and KalpnaGuleria. "Outbreak Trends of Coronavirus Disease–2019 in India: A Prediction." *Disaster medicine and public health preparedness* (2020): 1-6.

[12] Arti, M. K., and Kushagra Bhatnagar. "Modeling and Predictions for COVID 19 Spread in India." *ResearchGate, DOI: DOI* 10.

[13] Pandey, Gaurav, Poonam Chaudhary, Rajan Gupta, and Saibal Pal. "SEIR and Regression Model based COVID-19 outbreak predictions in India." *arXiv preprint arXiv:2004.00958* (2020).

[14] Received from https://www.covid19india.org/.

[15] Witten, Ian H. "Data mining with weka." *Department of Computer Science University of Waikato New Zealand* (2013).

[16] Kaur, Taranjot, Sukanta Sarkar, Sourangsu Chowdhury, Sudipta Kumar Sinha, Mohit Kumar Jolly, and ParthaSharathi Dutta. "Anticipating the novel coronavirus disease (COVID-19) pandemic." *medRxiv* (2020).

[17] Bhatia, Surbhi, Manisha Sharma, and Komal Kumar Bhatia. "A novel approach for crawling the opinions from world wide web." *International Journal of Information Retrieval Research (IJIRR)* 6, no. 2 (2016): 1-23.