Measurement: Sensors 25 (2023) 100585

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

COVID-19 outbreak data analysis and prediction

R. Anandan [a], T. Nalini [b,*], Shwetambari Chiwhane [c], M. Shanmuganathan [d], P. Radhakrishnan [e]

[a] Dept of C.S.E, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Pallavaram, Chennai, 600117, Tamil Nadu, India
[b] Dept of C.S.E, Dr.M.G.R.Educational and Research Institute, Maduravoyal, Chennai, 600095, Tamil Nadu, India
[c] Dept of C.S.E, Symbiosis Institute of Technology, Symbiosis International University, Lavale, Pune, India
[d] Dept of C.S.E, Panimalar Engineering College, Chennai, 600123, Tamil Nadu, India
[e] School of CS and AI, S.R.University, Warangal, Telangana, India

ARTICLE INFO

Keywords:
Regression
Linear-regression
Covid-19
Data analysis
MERS
SARS

ABSTRACT

Covid-19 is a novel pandemic disease with no potential vaccine treatment or medicine, the world is facing currently as of now. The death toll has increased to several lakhs and recovery rate is comparatively very less, was initially spotted in Wuhan (China). This spreads through close contact with people and socializing. The number of infected people varies with different parts of the world In our particular country India we are going through the lock down period which is the only vaccine to promote "social distancing" The hurdle arose due to the widespread of corona is major economy loss in combo with innocent lives. In this manuscript, we are visualizing the dataset which is publicly available to map, differentiate and separate the data in order to segregate the places that are most prone and perform basic regression to identify and predict the increasability of the counts from the dataset.

1. Introduction

Covid-19 which is also called the Corona Virus is a pandemic that the world is facing currently. It has spread globally since its first identification. The virus typically spreads among people with close contact and socializing. The first Coronavirus outbreak occurred in Wuhan, China around the time of December 2019. Later there were cases that appeared in Thailand [1]. Now, this has transmitted to more than 70 countries around the world. WHO confirmed 76000 cases of COVID-19 worldwide as of 30 January 2020 [2]. Coronavirus is a disease that ranges from normal symptoms like cold, cough, Middle East Respiratory Syndrome (MERS), Pneumonia to Severe Acute Respiratory Syndrome (SARS) though the symptoms are undetected until a few days [3]. If the symptoms get worse with the affected person losing immunity then this may lead to death [3]. Now the doctors say even if the coronavirus enters the human body one cannot find it out that easily because it is going symptom-less and one person can infect tons of people without even getting identified as corona positive. If the person is detected corona positive then it is isolated for fifteen days which is termed as incubation period and more until the person is completely cured, worldwide around 2,322,320 people are affected by this novel virus [4]. The death rates have been massively increasing throughout 210 countries. The demographics vary with different parts of the country and continents. There is no vaccination as of such in this current situation other than social distancing. The precautions and control measures are taken in such ways they are listed as follows. This social distancing in the world going on is in the form of lockdown in the countries where the colleges, universities, public sectors, private sectors, restaurants, hotels, and public gatherings everything is all closed the people themselves are locked down in their respective homes. Only hospitals are opened doctors and medical staff is allowed to work, the media persons but that too 1 m of distance to be maintained is the conditions applied while interviewing, the slogan of "stay home stay safe "is followed. The precautions taken are washing hands, using alcohol-based sanitizers, and not frequently touching the face. Basically, the negative impact of lockdown and coronavirus will certainly fall on people who are losing their lives and the economy of the world will find a huge fat loss, the victim of this virus is the poor, the lower middle class who tries to earn their daily bread by daily/monthly wages who is now left with no money to survive in the lockdown.

So, this kind of drastic situation should be analyzed and drastic effects should be predicted also the counts should be measured wherein the role of this manuscript takes place in the prediction of the active cases and the total number of cases. (see Table 2 and 3, Figs. 1–6)

* Corresponding author.
E-mail addresses: anandan.se@velsuniv.ac.in (R. Anandan), nalini.cse@drmgrdu.ac.in (T. Nalini), shwetambari.chiwhane@sitpune.edu.in (S. Chiwhane), shanmail2k@gmail.com (M. Shanmuganathan), rksiva13@gmail.com (P. Radhakrishnan).

https://doi.org/10.1016/j.measen.2022.100585
Received 5 October 2022; Received in revised form 2 November 2022; Accepted 20 November 2022
Available online 5 December 2022

Table 1
Reported cases as of 27th March 2020.

S. No	Name of State/UT	Total Confirmed Cases (Indian National)	Total Confirmed Cases (Foreign National)	Cured	Death
1.	Andhra Pradesh	12	0	1	0
2.	Chattisgarh	6	0	0	0
3.	Delhi	38	1	6	1
4.	Gujarat	43	0	0	3
5.	Harayana	16	14	11	0
6.	Himachal Pradesh	4	0	0	1
7.	Karnataka	20	0	3	2
8.	Kerala	131	7	11	0
9.	Madhya Pradesh	23	0	0	1
10.	Maharashtra	144	3	15	4
11.	Odisha	3	0	0	0
12.	Puducherry	1	0	0	0
13.	Punjab	29	0	0	1
14.	Rajasthan	41	2	3	0
15.	Tamil Nadu	32	3	1	1
16.	Telangana	34	11	1	0
17.	Chandigarh	7	0	0	0
18.	Jammu and Kashmir	18	0	1	1
19.	Ladakh	13	0	0	0
20.	Uttar Pradesh	42	1	11	0
21.	Uttarkhand	4	0	0	0
22.	West Bengal	11	0	0	1
23.	Bihar	7	0	0	1
24.	Mizoram	1	0	0	1
25.	Goa	6	0	0	0
26.	Manipur	1	0	0	0

2. Literature review

The Arogya Setu App is proposed by the Indian government where in the self-health analysis is taken into consideration and the citizens of India, the users are given enough knowledge, tips and treatments besides the precautions to deal with coronavirus. The Britain government has proposed an application wherein the smartphone tracking system is used wherein a person whom you crossed can be detected whether he/she is corona positive or not and we can take safety measures on that analysis. The application called COVID-19 Global Case Tracker is an interactive map provides people with the most up-to-date information on the coronavirus pandemic to capture all confirmed COVID-19 cases, fatalities, affected zones, regions, and recoveries. Also, a web application called the Worldomater gives pandemic information from around the world.

3. Proposed system

This manuscript is basically based on the real-time scenario in which data science participates where in data analysis plays a major role it is described in the numerical form, represented graphically it helps to read the data and also prediction process takes place for the further queries in future which are done by time series analysis we can also predict the deaths so the corona cases will be detected and will be reported in the numeric format then will set out in the graphical format with active cases and total cases can be recognized.

For example, as per Table 1, a state in India says Maharashtra has a certain number of cases of corona it will give the exact figure and also can compare with all other states, it will display the recoveries of patients, the upcoming statistics by moving the cursor across the map presented not only this but also will evaluate regions as such where the number of cases is relatively high and where it is low, it will also give the outline of the total number of cases and highlight the prospective areas, differentiate the hotspot high alert zone areas. This work can be very useful for the general public and also the government because it will give the ratio proportion and instead of manual/database records this project will be more valuable, it can keep the records transparent to the public and can be used by general public awareness method wherein we can spread awareness on the ground levels as well. In the jupyter notebook using python, Pandas library is used for playing with the data, Matplotlib is used for plotting graphs, the style used in the graph is ggplot, Plotly and Plotly Express is also used, folium assists us to use the map to explain in which zone the cases are appearing or reducing or about the recoveries and deaths, the data frames are set up. The linear regression supports the dataset by performing the activities and stating the relationships like giving the proper analysis about the variables involved such as the independent and dependent, it gives the statistics also

Table 2
Reported cases as of 27th March 2020 with Total and Active Cases.

S.No	Name of State/UT	Total Confirmed Cases (Indian National)	Total Confirmed Cases (Foreign National)	Cured	Death	Total Cases	Active Cases
1.	Andhra Pradesh	12	0	1	0	12	11
2.	Chattisgarh	6	0	0	0	6	6
3.	Delhi	38	1	6	1	39	32
4.	Gujarat	43	0	0	3	43	40
5.	Harayana	16	14	11	0	30	19
6.	Himachal Pradesh	4	0	0	1	4	3
7.	Karnataka	20	0	3	2	20	15
8.	Kerala	131	7	11	0	138	127
9.	Madhya Pradesh	23	0	0	1	23	22
10.	Maharashtra	144	3	15	4	147	128
11.	Odisha	3	0	0	0	3	3
12.	Puducherry	1	0	0	0	1	1
13.	Punjab	29	0	0	1	29	28
14.	Rajasthan	41	2	3	0	43	40
15.	Tamil Nadu	32	3	1	1	35	33
16.	Telangana	34	11	1	0	45	44
17.	Chandigarh	7	0	0	0	7	7
18.	Jammu and Kashmir	18	0	1	1	18	16
19.	Ladakh	13	0	0	0	13	13
20.	Uttar Pradesh	42	1	11	0	43	32
21.	Uttarkhand	4	0	0	0	4	4
22.	West Bengal	11	0	0	1	11	10
23.	Bihar	7	0	0	1	7	6
24.	Mizoram	1	0	0	1	1	1
25.	Goa	6	0	0	0	6	6
26.	Manipur	1	0	0	0	1	1

explains the influence in the change and peculiar idea about the increment, decrement, deaths in the corona cases also the active cases and the total number of cases in the country.

4. Methodology

4.1. Collecting the dataset

In this project, we have collected the Dataset from GitHub which is an open source platform. The data is then converted into a CSV (Comma Separated Values) file which is imported into the program to process and visualize the data.

4.2. Linear regression

Basic regression analysis is performed on the dataset where we determine the relations amidst a dependent variable and one or more independent variables depend on the given aspects [5]. We are using a set of statistical processes, to estimate the records of the Number of people deceased, active cases, amount of death, Number of people cured, Location of the deceased.

We are using Linear Regression because it serves the purpose to find which factor is more influencing change.

The regression result shows whether the relationship is valid. Simple linear regression is a kind of regression analysis in which the number of

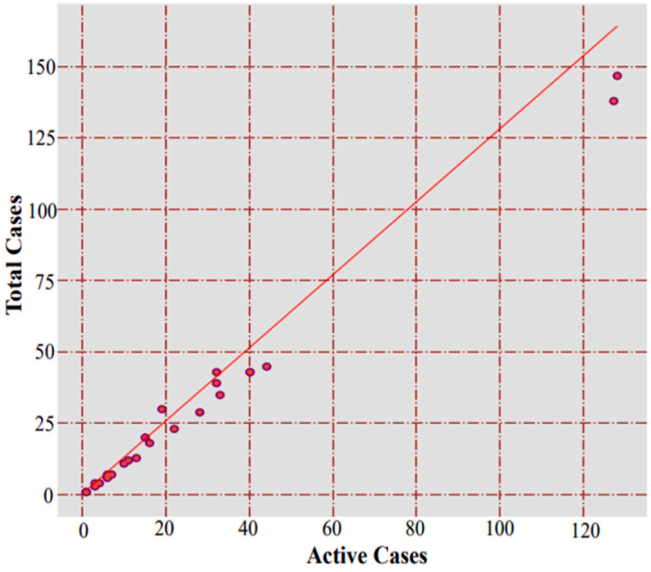

Fig. 1. Output of linear regression.

Table 3
Color Spectrum of Different Segments of Data.

S.No.	Name of State/UT	Total Confirmed Cases (Indian National)	Total Confirmed Cases (Foreign National)	Cured	Death	Total cases	Active Cases
1	Andhra Pradesh	12	0	1	0	12	11
2	Chhattisgarh	6	0	0	0	6	6
3	Delhi	38	1	6	1	39	32
4	Gujarat	43	0	0	3	43	40
5	Haryana	16	14	11	0	30	19
6	Himachal Pradesh	4	0	0	1	4	3
7	Karnataka	20	0	3	2	20	15
8	Kerala	131	7	11	0	138	
9	Madhya Pradesh	23	0	0	1	23	22
10	Maharashtra	144	3	15	4	147	128
11	Odisha	3	0	0	0	3	3
12	Puducherry	1	0	0	0	1	1
13	Punjab	29	0	0	1	29	28
14	Rajasthan	41	2	3	0	43	40
15	Tamil Nadu	32	3	1	1	35	33
16	Telangana	34	11	1	0	45	44
17	Chandigarh	7	0	0	0	7	7
18	Jammu and Kashmir	18	0	1	1	18	16
19	Ladakh	13	0	0	0	13	13
20	Uttar Pradesh	42	1	11	0	43	32
21	Uttarakhand	4	0	0	0	4	4
22	West Bengal	11	0	0	1	11	10
23	Bihar	7	0	0	1	7	6
24	Mizoram	1	0	0	0	1	1
25	Goa	6	0	0	0	6	6
26	Manipur	1	0	0	0	1	1

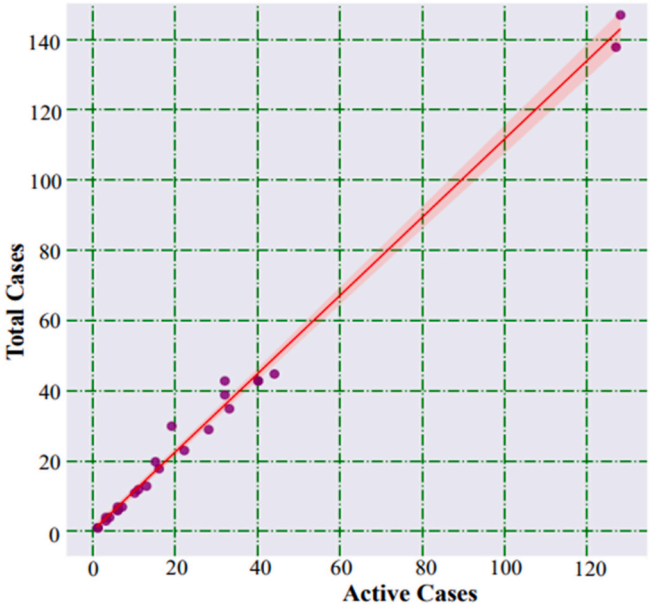

Fig. 2. Output of predicted values of Y.

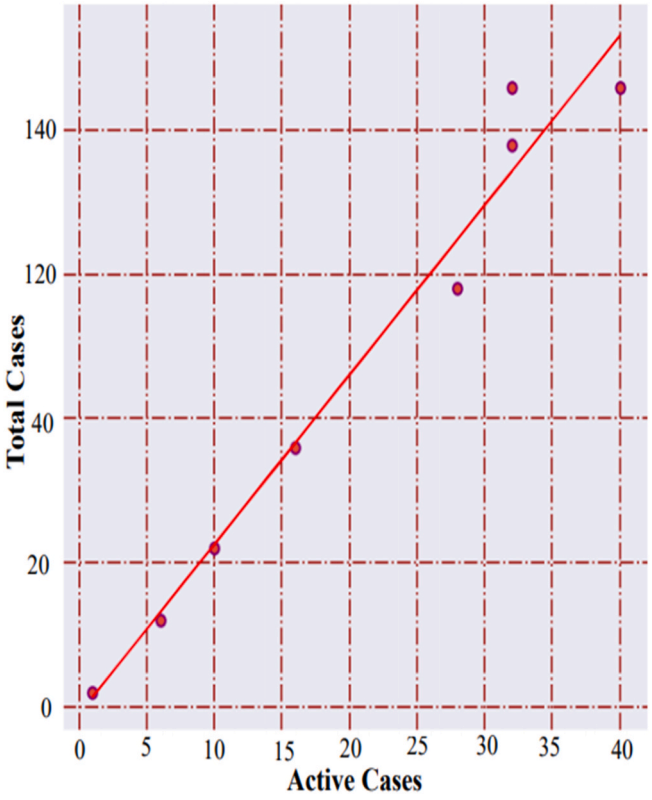

Fig. 3. Output of R-square model.

in dependent variables is one and there may be a linear an accord among the independent(x) and dependent(y) variable [8–10].

From the dataset, Dependent variable(X) would be total number of Cases and the Independent variable(Y) would be Active Cases.

The red highlighted line in the above graph is represented as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

From the above graph the shaded red part represents the predicted values of Y. Independent (active Cases/x) which could be one or more

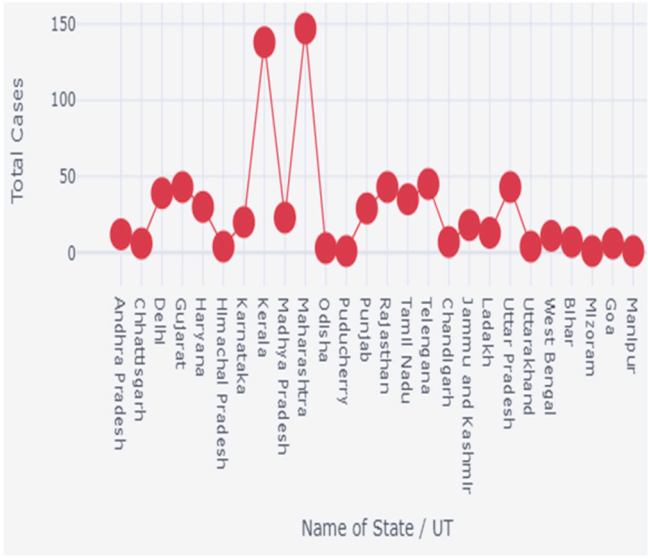

Fig. 4. Scatter plot.

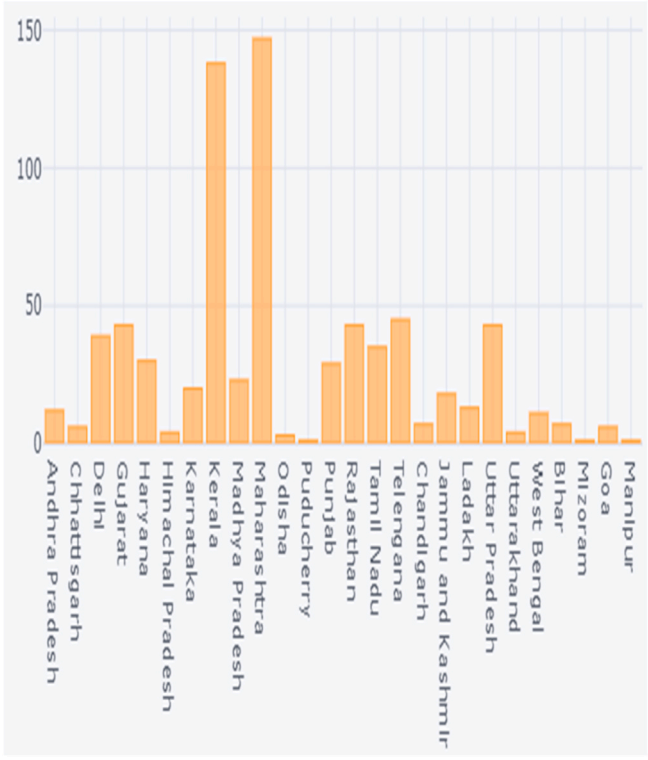

Fig. 5. Bar graph.

Dependent variable (Total cases/Y).

The steps to be followed for building a mathematical mod model are as follows:

- ➢ Find the Independent variable (x) and Reliant variable(y) and denote it as x & y.
- ➢ Evaluate the median of x and y.
- ➢ Determine the linear Regression equation as $Y = m\,x + c$.
- ➢ Find the slope m from the above equation as $m = \sum[x-x]\,[y-y]\,\sum[x-x]2$
- ➢ Find the constant c from the above equation.

- ➢ From the given values of m and c, determine the predicted values of y.
- ➢ Compare the distance between actual and predicted value.
- ➢ Find the goodness of fit using the R2 method [6].

4.3. R-squared(R2)

R-squared is a form of approach which is used to grade goodness-of-fit ranging for linear regression models. It is described as the proportion of deviation in the dependent or predicted variable (y) independent variable (x) using the best-fit line propagated by the regression analysis [7].

R-squared gives the proportion of the based dependent variable (y) variation that a linear model explains.

R^2 = Variance explained by the model / Total Variance

Residuals represent the distance amid the noticed value and fitted value.

R-squared is always between zero percent to a hundred percentage:

- ➢ Zero percentage is denoted while the model does not explain any of the fluctuations in the return variable around its mean. The mean of the dependent variable predicts the dependent variable and the regression model.
- ➢ A hundred percent intends that a model that explains all of the variations in the return variable i.e., the response around its mean.

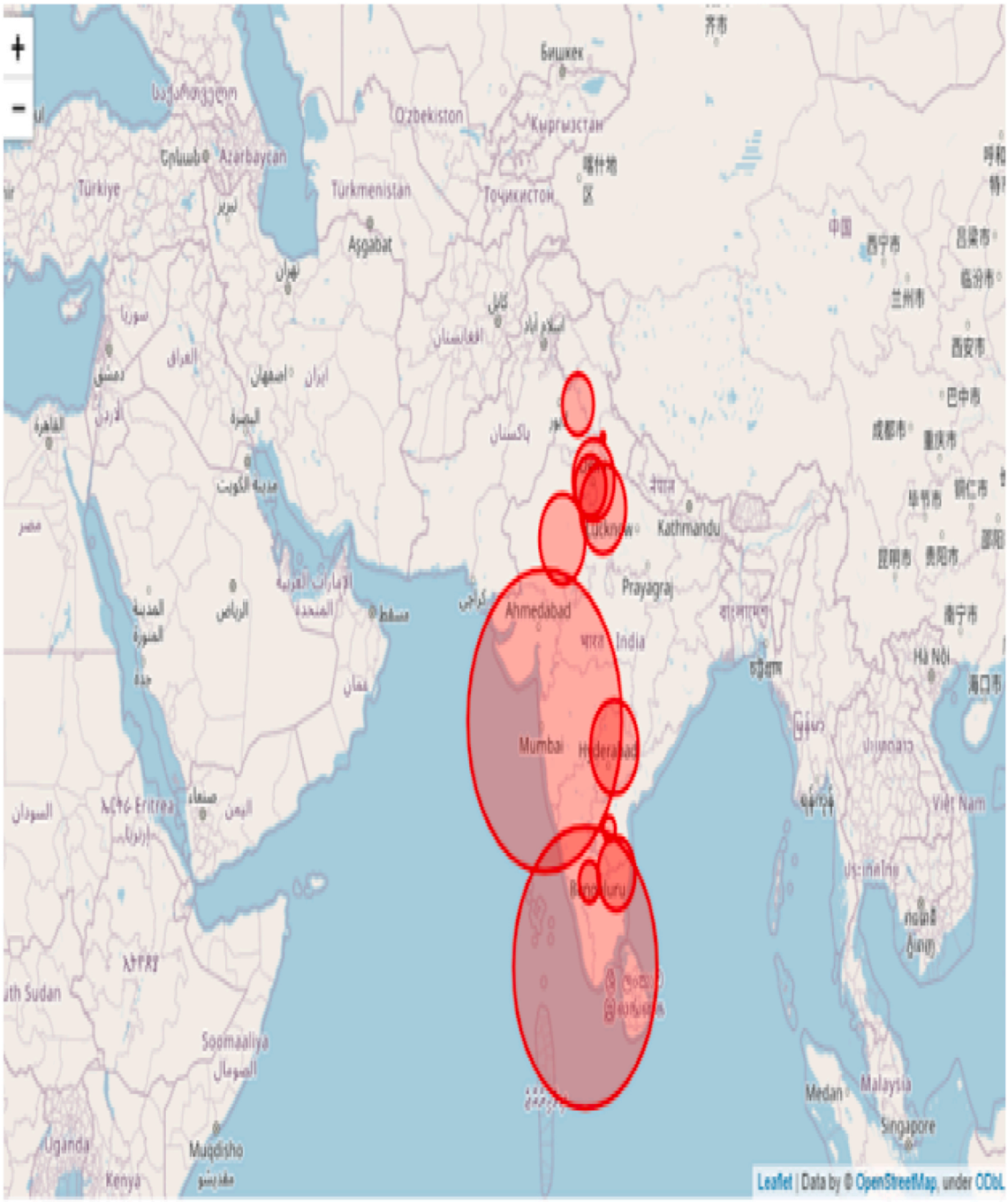

Fig. 6. World map view of the affected regions.

4.4. Heat map

A Heat map uses a wide color spectrum to denote different segments of data. The Darker the color spectrum the higher the cases and affected areas. The lighter color tones represent the average and less affected areas. The heat maps are highly used to visualize data, segregate into a more meaningful and understandable manner [11].

4.5. Scatter Plot

Scatter plots are typically used to visualize the dispersal of scattered dots or marks to the relationship of the variables which use the Cartesian coordinates to display the values i.e. the scatter gram.

The main purpose of a scatter chart is to show the type of relationship. Correlation, that exists between two sets of data or the variables [12,13].

4.6. Bar chart

Bar charts are generally used to visualize the comparisons between categories of data in the dataset. This could be displayed in two forms namely the vertical projection and the other is horizontal projection [14].

The World map view of the affected regions, wherein the darker tones i.e., the redness in the map defines more infected cases detected. Lighter tones represent commonly average or lower instances in that region.

5. Conclusion

This research presented current trends of COVID-19 outbreak till 27th March 2020. This process will be helpful in order to compare the situations based on the values. The graphical representations, plotting's give us a proper display of the affected areas, active proceedings and death rate. Further updates could be done to the system with UpToDate data additions which would follow the same analysis method. Tool used was jupyter notebook. Matplotlib is used for plotting graphs for visualization. Predicted the deaths of different states because of COVID-19 using Time Series Analysis. Majority of death rate was 2020 crude(per 1000 people), we used 2020 dataset.

CRediT authorship contribution statement

R. Anandan: Conceptualization, T. Nalini: Writing, Shwetambari Chiwhane, M. Shanmuganathan-review and editing, P. Radhakrishnan–editing and Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Hui, et al., The continuing 2019-nCoV epidemic threat of novel coronavirus to global health - the latest 2019 novel coronavirus outbreak in Wuhan, China, Int. J. Infect. Dis. 91 (2020) 264–266.
[2] World Health Organization WHO, Coronavirus Disease 2019 (COVID-19) Situation Report - 35, WHO, 2020.
[3] WHO, Advice for Public, WHO Int., 2020 [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public. (Accessed 27 February 2020).
[4] World Health Organization, Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19), World Health Organization, 2020.
[5] https://www.geeksforgeeks.org/ml-linear-regression/.
[6] K. Kumari, S. Yadav, Linear regression analysis study, J Pract Cardiovasc Sci [serial online] 4 (2018) 33–36 [cited 2020 May 3].
[7] https://statisticsbyjim.com/regression/interpret-r-squared-regression/.
[8] https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a.
[9] https://medium.com/datadriveninvestor/machine-learning-algorithms-linear-regression-f89ab64ac490.
[10] https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.
[11] https://www.tutorialspoint.com/python_data_science/python_heat_maps.htm.
[12] https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7.
[13] https://www.marsja.se/how-to-make-a-scatter-plot-in-python-using-seaborn/.
[14] https://medium.com/python-pandemonium/data-visualization-in-python-bar-graph-in-matplotlib-f1738602e9c4.