ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.7, 4579–4587
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.7.26 Spec. Issue on RDESTM-2020

EXPLORATION ON COVID-19 DATA IN INDIA USING MACHINE LEARNING FOR PREDICTION OF INFECTED AND DEATH CASES

P. NANCY[1], S. SRIDHAR, R. AKILADEVI, AND V. SUDHA

ABSTRACT. The name that is prevalent everywhere today throughout the world is COVID-19. COVID -19 is an infectious disease that affects the lungs and transmitted through droplets generated when an infected person coughs, sneezes, or exhales. It is caused by virus named coronavirus and originated from China in December 19. Various factors have been identified for spread of corona, but none has been proved. Hence we made an attempt in identifying correlation between temperature and various case of corona in states of India. Machine learning techniques have been adopted to find interesting information. The data set for this analysis is collected from Kaggle. The novelty of the work is that temperature data has been included in the original data to explore future inclinations. We performed liner regression model for prediction of death, confirmed cases and recovery. The research findings show that the effect of temperature in the states is different for different cases(confirmed, cured, death) of COVID-19.

1. INTRODUCTION

Research has proved that seven types of virus under corona family affect the humans though many come under corona family [1]. COVID-19 is a pandemic disease caused by corona virus with a variety of symptoms varying from mild cold, cough and breathlessness and might end up fatal too. There is a strong

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 62J05, 68T09, 60G25L.

Key words and phrases. COVID-19, India, Machine Learning, Cured, Death, Confirmed.

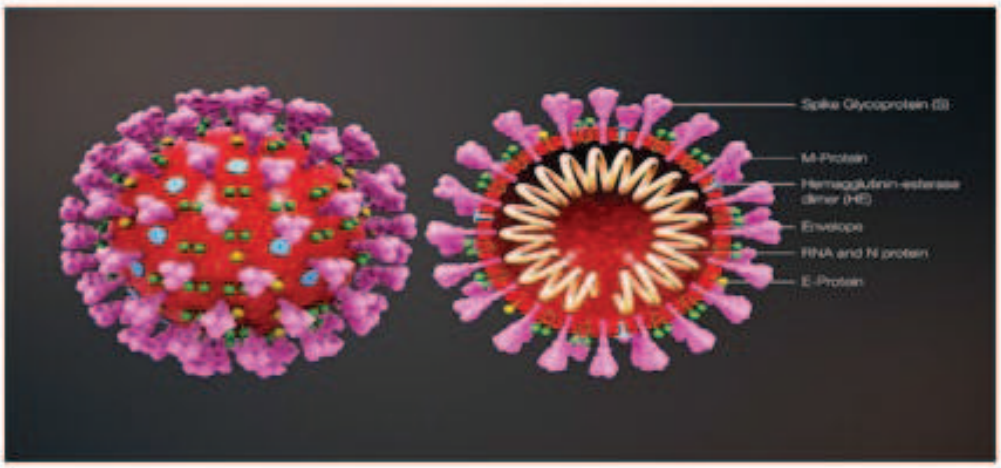

FIGURE 1. THE STRUCTURE OF CORONA VIRUS.

belief that the virus causing the disease is transmitted from animals like bats to humans [1] [2]. This disease was first identified in Wuhan city of China in December 2019 [3]. The people infected with this disease might experience mild to moderate respiratory illness and might recover without even undergoing any special treatment. However, elderly people and people with medical problems like cardiovascular disease, Diabetes. Chronic respiratory disease and cancer are more prone to develop severe illness. The spread of Covid-19 virus predominantly occurs by droplets of saliva or discharges from the Nose, if an infected person coughs or sneezes. As of now, there is no preventive vaccination or medical treatment for this disease. The World Health Organization strives hard for stopping the spread of epidemic and providing care for the affected people. The structure of Corona virus is shown in Figure 1 [4].

Every nation urged with different strategies to minimize the spread of this pandemic starting from Quarantine, isolation, usage of masks, curfew, social distancing, sanitizing, hand washing. The above said strategies were effectively insisted by Government of India. When a person is identified to be infected, He or She immediately given special focus for proper treatment in hospitals and the primary contacts of the infected person was traced and quarantined for 14 to 28 days. First case of Covid-19 in India was reported on January 30th, 2020

in Thrissur, Kerala. AS on 9th July 2020, the No. of Confirmed cases were 7, 67,296, Active cases are 2, 69,790 and recovered are 4, 76,377 and Death are 21,129 [5]. The Government divided entire nation into 3 Zones, Viz. Green, Red and Orange. The Districts without any confirmed cased or new cases in last 21 days fall under Green Zone, The Districts with fewer cases fall under Orange Zone (Non-Hot Spots), the districts with high doubling rate and higher number of active cases fall under Red Zone-Hot Spots. Worldwide spread of this Pandemic throws a big challenge to the Global public health community. The correct detection analysis becomes essential to overcome this circumstance. In spite of many mythologies that have been used extensively for disease diagnosis, prognosis and analysis. Machine Learning Methodologies shall be preferred for effective analysis and results. In this research, we try to extract the correlation between temperature and the different cases-Confirmed, Cured and death. The data set utilized for this research is obtained from KAGGLE [6].

2. METHODS AND MATERIALS

Dataset is very essential for analysis. In this research, we have collected the data Kaggle repository. This data covers the details of corona infected persons in different states of India. Since the research focuses on finding out the correlation of temperature related to confirmed, cured and death cases of corona infected persons, we added a separate column of temperature details of all the various states starting from the month of January to July. Of the various machine learning techniques we adopted Linear Regression (LR) to find out the association of temperature and Covid infected cases. Linear regression is one of the simplest and easier algorithms in statistics and machine. It helps to derive the correlation between two numeric input variables x and y in the form of an equation. The algorithm is used to predict the output for set of values of y, which is done using the values of input x.

The equation obtained will allot a scale factor to every input value, called a coefficient and represented by the capital Greek letter Beta (B). An additional value is added, which in turn derives intercept or the bias coefficient [7].

$$y = B0 + B1*x \text{ [7]}$$

The details of COVID-19 cases in India under all categories – confirmed, cured, death and active are shown in Table 1.The entire analysis is done using Weka (a data mining tool).

Table 1: COVID-19 PANDEMIC IN INDIA BY STATE AND UNION TERRITORY [5]

State / Union Territory	Cases	Death	Recovered	Active
Andaman and Nicobar Islands	151	0	83	68
Andhra Pradesh	23,814	277	12,154	11,383
Arunachal Pradesh	302	2	120	180
Assam	14,032[b]	22	8,729	5,281
Bihar	13,944	115	9,816	4,013
Chandigarh	523	7	403	113
Chhattisgarh	3,675	15	2,903	757
Dadra and Nagar Haveli and Daman and Diu	411	0	189	222
Delhi	107,051	3,258	82,226	21,567
Goa	2,151	9	1,273	869
Gujarat	39,194	2,008	27,718	9,468
Haryana	19,369	287	14,510	4,572
Himachal Pradesh	1,140	11	846	283
Jammu and Kashmir	9,501	154	5,695	3,652
Jharkhand	3,246	23	2,208	1,015
Karnataka	31,105	486	12,833	17,786
Kerala	6,534	27[c]	3,708	2,799
Ladakh	1,055	1	915	139
Lakshadweep	0	0	0	0
Madhya Pradesh	16,341	634	12,232	3,475
Maharashtra	230,599	9,667	127,259	93,673
Manipur	1,450	0	799	651
Meghalaya	113	2	66	45
Mizoram	197	0	133	64

Nagaland	673	0	304	369
Odisha	11,201	52	7,407	3,742
Puducherry	1,151	14	584	553
Punjab	7,140	183	4,945	2,012
Rajasthan	22,563	491	17,070	5,002
Sikkim	134	0	72	62
Tamil Nadu	126,581	1,765	78,161	46,655
Telangana	30,946	331	18,192	12,423
Tripura	1,776	1	1,338	437
Uttar Pradesh	32,362	862	21,127	10,373
Uttarakhand	3,305	46	2,672	587
West Bengal	25,911	854	16,826	8,231

TABLE 2. DETAILS OF ATTRIBUTES USED IN KAGGLE DATASET

Attribute	Detail
Date	Date of data collected
Time	Time
State / Union Territory	Name of the State
Cured	Number of persons cured / recovered from COVID-19
Death	Number of persons died due to COVID-19
Confirmed	Number of persons identified to have COVID-19

The Kaggle dataset has all the attributes as shown in Table 2 except one attribute- temperature. The major reason behind the addition of temperature attribute is to find the correlation of temperature with different cases of COVID. The addition of temperature attribute will be favorable in exploring the predicted patterns.

Row Labels	Andhra Pradesh	Delhi	Gujarat	Haryana	Karnataka	Madhya Pradesh	Maharashtra	Rajasthan	Tamil Nadu	Telangana	Uttar Pradesh	West Bengal
March	40	97	73	40	83	47	216	74	74		101	26
April	1403	3439	4082	310	557	2660	9915	2438	2162		2203	758
May	3569	18549	16343	1923	2922	7891	65168	8617	21184		7445	5130
June	13891	85161	31938	14210	14295	13370	169883	17660	86224	15394	22828	17907
July	20019	100823	36772	17504	25317	15284	211987	20688	114978	25733	28636	22987

FIGURE 2. Confirmed Cases in 12 States From March To July.

Row Labe	Andhra Pradesh	Delhi	Gujarat	Haryana	Karnataka	Madhya Pradesh	Maharashtra	Rajasthan	Tamil Nadu	Telengana	Uttar Pradesh	West Bengal
March	0	2	6	0	3	3	9	0	1	1	0	2
April	31	56	197	3	21	130	432	51	27	26	39	22
May	60	416	1007	20	48	343	2197	193	160	77	201	309
June	180	2680	1827	232	226	564	7610	405	1141	156	672	653
July	239	3115	1960	276	401	617	9026	461	1571		809	779

FIGURE 3. Death Cases in 12 States From March To July.

Row Labels	Andhra Pradesh	Delhi	Gujarat	Haryana	Karnataka	Madhya Pradesh	Maharashtra	Rajasthan	Tamil Nadu	Telangana	Uttar Pradesh	West Bengal
March	1	6	3	21	5	0	39	3	4		14	0
April	321	1092	527	209	223	461	1593	768	1210		513	124
May	2289	8075	9230	971	997	4444	28081	5739	12000		4410	1970
June	6232	56235	23240	9502	7683	10199	88960	13618	47749	5582	15506	11719
July	8920	72088	26315	13335	10527	11579	115262	16278	66571	14781	19109	15235

hh

FIGURE 4. Cured Cases in 12 States From March To July.

3. EXPERIMENTAL RESULTS

The research started with the analysis of number of cases confirmed, cured and dead in every month starting from January to July. Data is pre-processed by organising the data month wise with all the counts of different cases accordingly. The states with maximum counts under each category – confirmed, cured and death were identified which include Andhra Pradesh, Delhi, Gujarat, Haryana, Karnataka, Madhya Pradesh, Maharashtra, Rajasthan, Tamil Nadu, Telangana, Uttar Pradesh and West Bengal. Since the count of confirmed, cured and death cases are very meagre in the month of January and February, it has not been

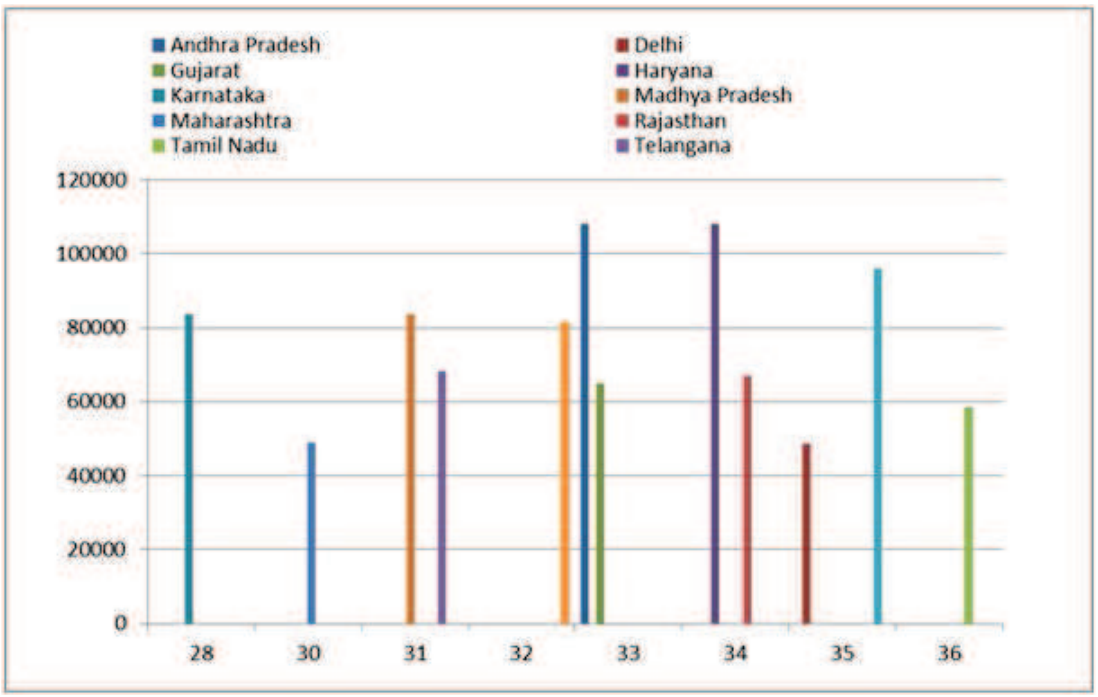

FIGURE 5. Predicted Confirmed Cases in 12 States Using Linear Regression.

utilised for the analysis. The details from March to July alone are considered. The number of persons confirmed, dead and cured due to COVID-19 from March to July in the identified 12 states are shown in Figure 2,3and 4.

The average temperature from March to July is calculated for all the 12 states and total count of confirmed, cured and death cases are taken as input for linear regression model and prediction were made for all the 3 cases related to temperature. 95 percentage confidence level is considered in the work [9]. In Figure 5 and 6 the X axis shows the temperature and Y axis shows the number of cases [9]. Figure 5 shows the predicted confirmed cases for the 12 states using LR. From the graph it can be inferred that the count of confirmed and death cases are going to be increased in some of the states and decreased in some states like Delhi, Maharashtra and Tamilnadu. Figure 6 shows the predicted death cases for the 12 states using LR. The graph can be interpreted that the cases are going to be increased in future as per the existing case data.

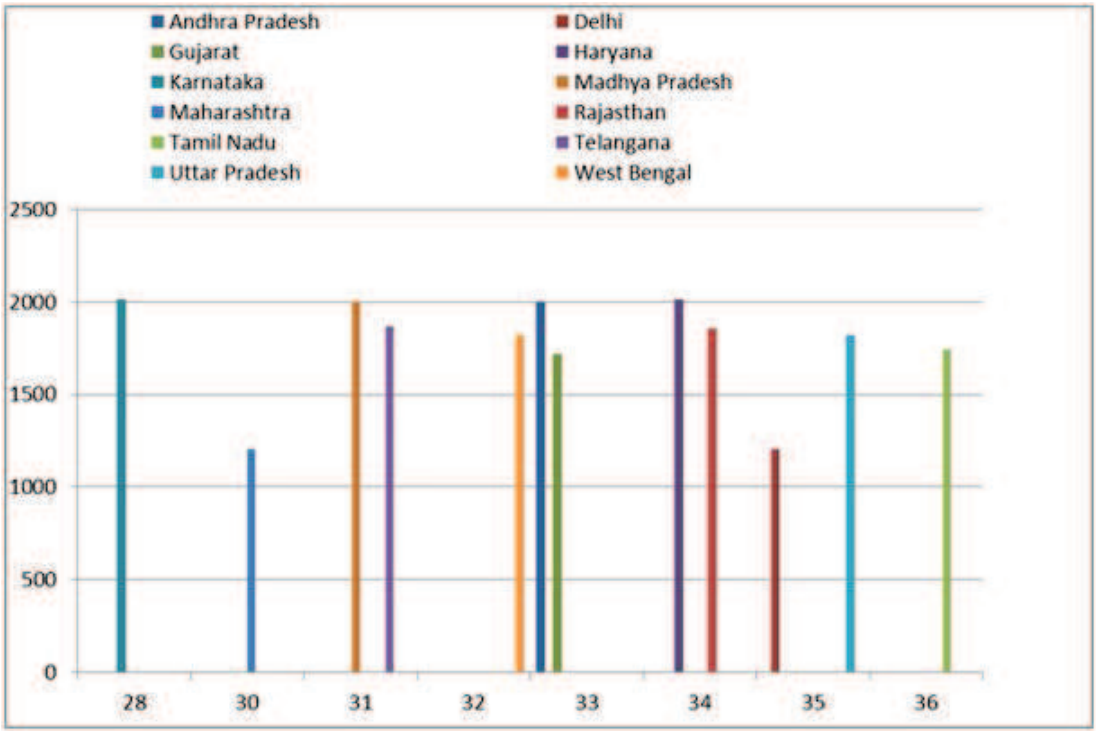

FIGURE 6. Predicted Death Cases in 12 States Using Linear Regression.

4. CONCLUSION AND FUTURE WORK

From the data analysis, it is clearly evident that temperature is the not the only important factor for spread of COVID-19. While utilising temperature attribute for the various cases of COVID-19 (confirmed, cured, death), it was observed that varied nature of trends exist for each state. This shows the necessity of additional attributes for effective prediction of patterns from the data. The future work in this aspect would be the inclusion of attributes like rainfall, humidity, hours of sunshine, wind speed in the original data and discover some useful patterns.

REFERENCES

[1] M. BILAL, M. NAZIR, R. PARRA-SALDIVAR, H,M. IQBAL: *2019- nCoV/COVID-19 - Approaches to Viral Vaccine Development and Preventive Measures*, J. Pur. App. Mic. **14**(1) (2020), 25–29.

[2] Y. FAN, K. ZHAO, Z.L. SHI, P. ZHOU: *Bat Coronaviruses in China*, Viruses, **11**(3) (2019), 210.

[3] B. GATES: *Responding to Covid-19-A Once-in-a-3. Century Pandemic?*, N. Eng. J. Med., **382**(18) (2020), 1677–1670.
[4] *Corona Virus Disease* https://en.wikipedia.org/wiki/Coronavirus_disease/3D-medical-animation-coronavirus-structure.eps
[5] *COVID-19 pandemic in India*: https://en.wikipedia.org/wiki/COVID-19-pandemic-in-India-Lockdown
[6] *KAGGLE Dataset*: https://www.kaggle.com/sudalairajkumar/covid19-in-india
[7] *Linear Regression:* https://machinelearningmastery.com/linear-regression-for-machine-learning/
[8] M. K. SIDDIQUI, R. MORALES-MENENDEZ, P. KUMAR, GUPTA, M.N. HAFIZ, IQBAL, F. HUSSAIN, K. KHATOON, S. AHMAD: *Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis*, J. Pur. App. Mic. **14** (Suppl 1) (2020), 1017-1024.
[9] R. SUJATH, J. MOY CHATTERJEE, A. ELLA HASSANIEN: *A Machine learning forecasting model for COVID-19 pandemic in India*, Sto. Env. Res. and Ris. Ass. **34** (2020), 959-972.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
RAJALAKSHMI ENGINEERING COLLEGE
CHENNAI, TAMIL NADU, INDIA.
Email address: nancy.p@rajalakshmi.edu.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SAVEETHA SCHOOL OF ENGINEERING
SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES
CHENNAI, TAMIL NADU, INDIA.
Email address: 007sridol@gmail.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
RAJALAKSHMI ENGINEERING COLLEGE
CHENNAI, TAMIL NADU, INDIA.
Email address: akiladevi.r@rajalakshmi.edu.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
KUMARAGURU COLLEGE OF TECHNOLOGY
TAMIL NADU, INDIA.
Email address: sudhavelvizhi@gmail.com