Annals of Data Science
https://doi.org/10.1007/s40745-022-00404-w

Monitoring COVID-19 Cases and Vaccination in Indian States and Union Territories Using Unsupervised Machine Learning Algorithm

S. Chakraborty[1]

Received: 10 July 2021 / Revised: 20 October 2021 / Accepted: 30 October 2021

Abstract
The worldwide spread of the novel coronavirus originating from Wuhan, China led to an ongoing pandemic as COVID-19. The disease being a contagion transmitted rapidly in India through the people having travel histories to the affected countries, and their contacts that tested positive. Millions of people across all states and union territories (UT) were affected leading to serious respiratory illness and deaths. In the present study, two unsupervised clustering algorithms namely k-means clustering and hierarchical agglomerative clustering are applied on the COVID-19 dataset in order to group the Indian states/UTs based on the pandemic effect and the vaccination program from the period of March, 2020 to early June, 2021. The aim of the study is to observe the plight of each state and UT of India combating the novel coronavirus infection and to monitor their vaccination status. The research study will be helpful to the government and to the frontline workers coping to restrict the transmission of the virus in India. Also, the results of the study will provide a source of information for future research regarding the COVID-19 pandemic in India.

Keywords COVID-19 · Corona virus · Pandemic · Data analysis · Clustering · k-means clustering · Hierarchical clustering · Agglomerative clustering

1 Introduction

The COVID-19 pandemic [1–4] originating from Wuhan, China, in December 2019 is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The cause of the respiratory disease was confirmed by WHO [4] as a novel coronavirus on 12th January 2020. The variable symptoms observed in the COVID-19

✉ S. Chakraborty
chakrabartysonali@gmail.com

[1] Department of Mathematical and Computational Sciences, National Institute of Technology Karnataka, Mangalore 575025, India

Published online: 04 May 2022

patients are fever, cough, fatigue, breathing difficulties, loss of smell and taste. Severe consequences were observed in the patients such as cytokine storms, multi-organ failure, septic shock, blood clots, damage to the lungs and heart.

The transmission increased rapidly in all countries through the people having travel histories to the affected countries and their contacts. In India, the first case of COVID-19 was reported in Kerala on 30th January 2020, which rose to three cases by 3rd February. After a gap of almost one month, 22 new cases were reported on 4th March 2020. Eventually, the transmission rate grew in March 2020, and the first COVID-19 fatality of India was reported on 12th March of a 76-year-old man, who returned from Saudi Arabia. The rapid spread of the disease led to a nationwide lockdown for 21 days starting from 25th March, 2020 which was further extended by 14 days and thereafter by further two weeks with convinced relaxations. Considering the population of India [3] of about 121,05,69,573, controlling the extensive consequences of the pandemic was quite challenging for the administrative officials. Initially, the authorities decided to test only those people who returned from high-risk countries or who came in contact with the positive cases. Due to the substantial increase in the number of cases, later the government decided to test the people with pneumonia cases, irrespective of travel or contact history. With strict restrictions and guidelines imposed by the government, the number of cases decreased to 9000 per day by February, 2021.

A second wave was observed throughout India from early April, 2021and by the end of the month over 4 lakh cases and more than 3500 deaths were reported in a day. India began its vaccination program from 16th January, 2021with two DCGI (Drug Controller General of India) approved drugs. The frontline workers i.e., doctors, nurses, hospital staff and policemen were the first one to receive doses of the vaccine. Thereafter, the vaccination drive was extended to all the residents over the age of 45 and later for all residents over the age of 18.

In the present study, two unsupervised clustering algorithms [5–8]; namely k-means clustering and hierarchical agglomerative clustering is applied on the Indian states/UTs to group them on the basis of their demographic characteristics, the number of confirmed and death cases due to COVID-19 and their vaccination status. The use of data science in monitoring COVID-19 cases and vaccination status helps in gaining insight from the dataset and extracting meaningful information which can be further used for predicting future patterns and behaviours. The popular unsupervised clustering algorithms are used due to the fact the available COVID-19 data set for India is untagged. The number of clusters to be formed is not known as so it is desirable that the clustering algorithm will divide the dataset into groups based on their similarities. Wide-ranging research is being carried out about the COVID-19 pandemic; therefore, a brief review is presented in Sect. 2 from the available limited literatures. Sections 3.3 and 3.4 discusses the COVID-19 confirmed and death situation in India and each state and UT of India respectively. The vaccination status for India and its states/UT considering their population is discussed in Sects. 3.5 and 3.6. Section 4 performs cluster analysis on the COVID-19 dataset using k-means and hierarchical agglomerative clustering algorithms. The concluding remarks and the future scope of the research are discussed in Sect. 5.

2 Review of Literatures

Extensive research is being carried out both pathologically and statistically on the COVID-19 dataset across the world in order to observe the trend of infection transmission and to combat the spread of novel corona virus. In India as on 8th June, 2021, more than two crore people got infected with novel corona virus and approximately 3.5 lakh people succumbed to COVID-19 [9]. As on 8th June, 2021, the total number of people affected with coronavirus across the world is more than 18 crore and more than 39 lakh deaths are reported [4]. Apart from losing precious lives, the pandemic has a severe impact on the Indian economy and led to a negative growth rate for the first time in decades.

The pathological research aims to study the evolution, replication, pathogenesis [10] the transmission trend of the novel corona virus [11], its clinical features, diagnosis, treatment [12], and to observe the impact of the pandemic based on the parameters such as air temperature, relative humidity [13], age and gender. To perform statistical research on the COVID-19 dataset, various statistical models are being used by the researchers [14–21] and artificial intelligence techniques [22] are being suggested for predicting the further spread of the pandemic. Gondauri et al. [23] uses BAILEY's model to study and analyse the cases based on corona virus spread in different countries. Based on the experimental results, the author concluded the state of the virus spread and recovery up to 30th March, 2020. A spline-based time series with Bayesian model is used by Kumar et al. [24] to identify the transmission stages of COVID-19 infection in India. The Susceptible-Exposed-Infectious-Recovered (SEIR) model is used by researchers Pai et al. [25] to forecast the active COVID-19 cases in India considering the effect of nationwide lockdown and the possible inflation in the active cases after unlocking the nation.

3 Data Analysis

3.1 Analysis Domain

The research data included in the present study is for India which is the second most populated country in the world having 28 states and 8 union territories. The total population [3] of the country is 121,05,69,573 with a density of 382/km^2. The rural and the urban population of India is 69% and 21% respectively. The states and UTs used in the study are listed alphabetically and their abbreviations used are: Andaman and Nicobar (AN), Andhra Pradesh (AP), Arunachal Pradesh (AR), Assam (AS), Bihar (BH), Chandigarh (CH), Chhattisgarh (CG), Dadra and Nagar Haveli (DNH), Delhi (DL), Goa (GA), Gujarat (GJ), Haryana (HR), Himachal Pradesh (HP), Jammu and Kashmir (JK), Jharkhand (JH), Karnataka (KA), Kerala (KL), Ladakh (LD), Lakshadweep (LK), Madhya Pradesh (MP), Maharashtra (MH), Manipur (MN), Meghalaya (MG), Mizoram (MZ), Nagaland (NG), Odisha (OD), Pondicherry (PD), Punjab (PJ), Rajasthan (RJ), Sikkim (SK), Tamil Nadu (TN), Telangana (TL), Tripura (TR), Uttar Pradesh (UP), Uttarakhand (UT), West Bengal (WB).

3.2 Data Collection and Tools Used for Analysis

The COVID-19 dataset for India and the demographic details of each state and UT used in this study is extracted from the official website administered by the government of India [1, 2, 9]. The analysis includes approximately 16,000 records in comma separated values (CSV) format containing day wise information of the number of confirmed cases, active cases, cured cases and the number of deaths from March, 2020 to 8th June, 2021. The vaccination details contain the total number of people vaccinated for the first and the second dose starting from 16th January, 2021 in all states and UTs of the country for different age groups. The computations are performed using Microsoft Excel 2010 and RStudio Desktop 1.3.1093 is used for implementing the clustering algorithms.

3.3 Analysis of COVID-19 Confirmed and Death Cases of India

Table 1 depicts the total COVID-19 confirmed cases and deaths in India from March, 2020 to 8th June, 2021. A graphical representation of the same is depicted in Figs. 1 and 2 respectively.

The trend of covid cases observed in Table 1 and Figs. 1 and 2 depict that from March, 2020 the number of covid cases started rising in India. During September, 2020, the covid cases were at peak and thereafter they started declining towards February, 2021. This period of twelve months can be considered as the first wave of the pandemic

Table 1 COVID-19 confirmed cases and deaths in India from March, 2020 to 8th June, 2021

Month, Year	Total confirmed cases	Total deaths
March, 2020	1397	35
April, 2020	31,653	1039
May, 2020	1,49,093	4090
June, 2020	3,84,697	11,729
July, 2020	10,72,030	18,854
August, 2020	19,82,375	28,722
September, 2020	26,04,518	33,028
October, 2020	19,11,356	24,144
November, 2020	12,94,572	15,498
December, 2020	8,34,983	11,599
January, 2021	4,79,509	5536
February, 2021	3,33,796	2664
March, 2021	10,69,356	5530
April, 2021	66,13,641	45,862
May, 2021	92,84,558	1,20,770
June, 2021	12,27,289	33,979
Total	2,92,74,823	3,62,661

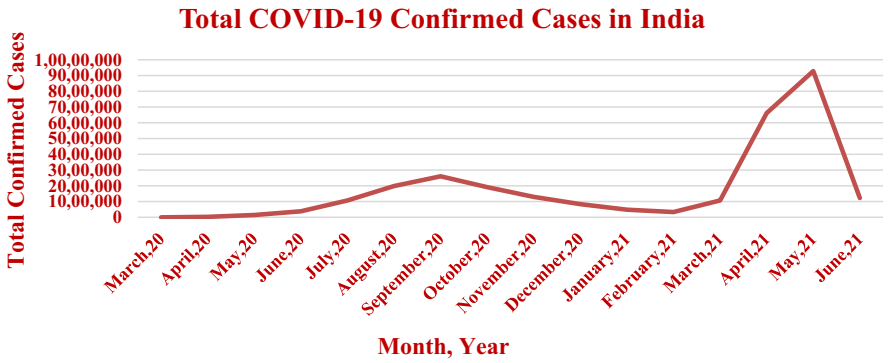

Fig. 1 Total confirmed COVID-19 cases in India from March, 2020 to 8th June, 2021

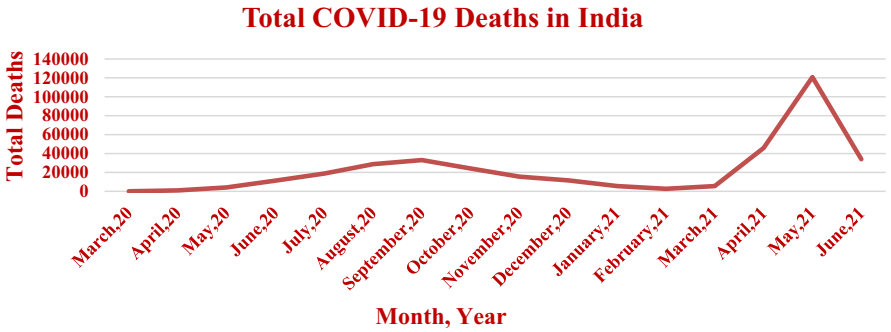

Fig. 2 Total COVID-19 deaths in India from March, 2020 to 8th June, 2021

in India. Again, from March, 2021 the number of covid cases started increasing and during May, 2021 the number of cases were maximum. Eventually, the cases started decreasing by the onset of June, 2021. This time period of four months can be considered as the second wave of the pandemic in India. Although the duration of the first and the second wave are not same, the total number of cases and the peak observed during the second wave with 92,84,558 confirmed cases in the month of May, 2021 is much higher than that observed during the first wave. The second wave in India was more fatal as compared to the first wave.

3.4 Analysis of COVID-19 Confirmed Cases and Deaths in the States and UTs of India

Table 2 depicts the total COVID-19 confirmed cases and deaths in each state and UT of India during the first and the second wave. A trend of the same is depicted in Figs. 3, 4, 5 and 6.

It is observed that among all the states and UTs, Maharashtra reported the highest number of COVID-19 cases and deaths both during the first and the second wave. All

Table 2 COVID-19 confirmed cases and deaths in Indian states and UTs during first and second wave

State/UT	First wave		Second wave	
	Total confirmed cases	Total deaths	Total confirmed cases	Total deaths
AN	5018	62	2113	61
AP	8,89,799	7169	8,73,412	4383
AR	16,836	56	12,860	69
AS	2,17,527	1092	2,21,219	2603
BH	2,62,509	1541	4,51,370	3883
CH	21,719	351	38,988	423
CG	3,12,419	3833	6,69,441	9410
DNH	3406	2	7013	2
DL	6,39,092	10,909	7,90,383	13,718
GA	54,932	794	1,04,879	2046
GJ	2,69,482	4409	5,47,530	5535
HR	2,70,610	3047	4,92,321	5704
HP	58,598	995	1,37,157	2320
JK	1,26,383	1956	1,75,084	2134
JH	1,19,905	1088	2,21,671	3972
KA	9,50,730	12,326	17,56,751	19,594
KE	10,56,149	4182	15,86,246	5975
LD	9818	130	9440	65
LK	359	1	8416	40
MP	2,61,403	3863	5,24,364	4506
MH	21,46,777	52,092	36,95,223	48,378
MN	29,271	373	26,557	523
MG	13,961	148	25,195	521
MZ	4423	10	9573	45
NG	12,199	91	10,719	338
OD	3,37,104	1915	4,82,110	1120
PD	39,717	668	69,844	970
PJ	1,81,597	5825	3,99,232	9335
RJ	3,20,180	2787	6,26,795	5900
SK	6137	135	11,033	138
TN	8,51,063	12,493	14,05,618	14,863
TL	2,98,807	1634	2,98,006	1792
TR	33,417	391	22,752	181
UP	6,03,427	8725	10,95,656	12,608
UT	97,031	1692	2,37,388	5039
WB	5,74,926	10,266	8,57,093	6096

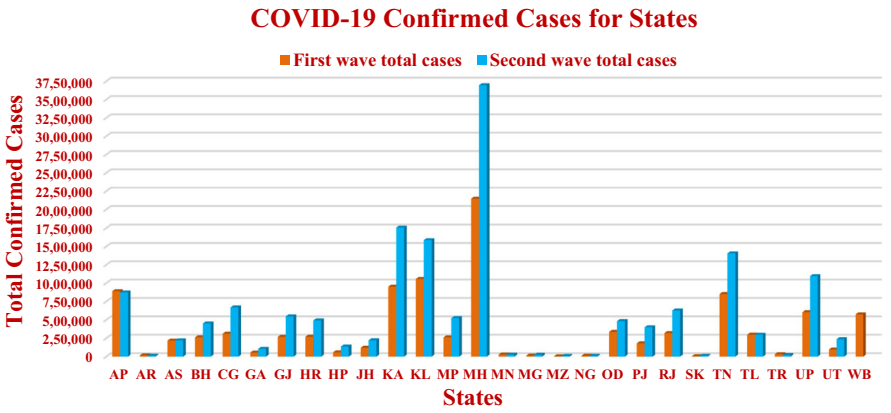

Fig. 3 COVID-19 confirmed cases in states of India during first and second wave

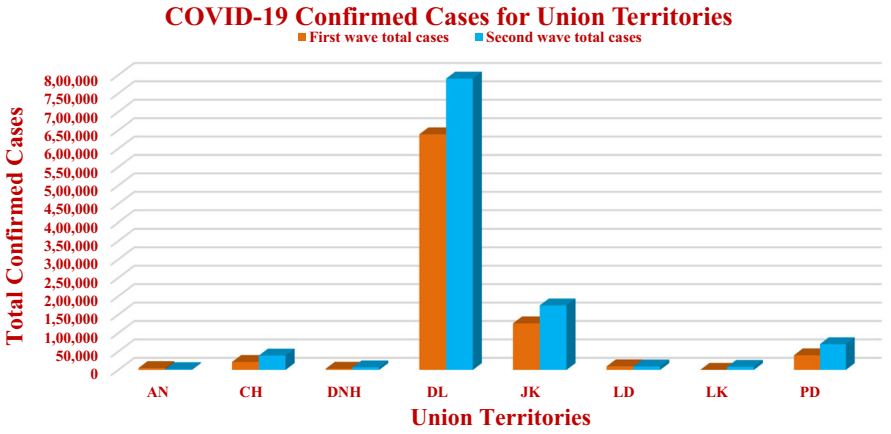

Fig. 4 COVID-19 confirmed cases in union territories of India during first and second wave

other states/UTs except Andaman and Nicobar (AN), Andhra Pradesh (AP), Arunachal Pradesh (AR), Ladakh (LD), Manipur (MN), Nagaland (NG) and Telangana (TL) showed a greater number of cases during the second wave as compared to the first wave. Lakshadweep reported its first COVID positive case in January, 2021. States Andhra Pradesh, Delhi, Karnataka, Kerala, Tamil Nadu, Uttar Pradesh and West Bengal reported more than 5 lakh cases during both the waves.

3.5 COVID-19 Vaccination Status in India

Looking into the severity of the second wave, the foremost priority of the government is to speed up the vaccination process among the residents. No vaccines were available almost during the first wave. The vaccination process started in India from 16th

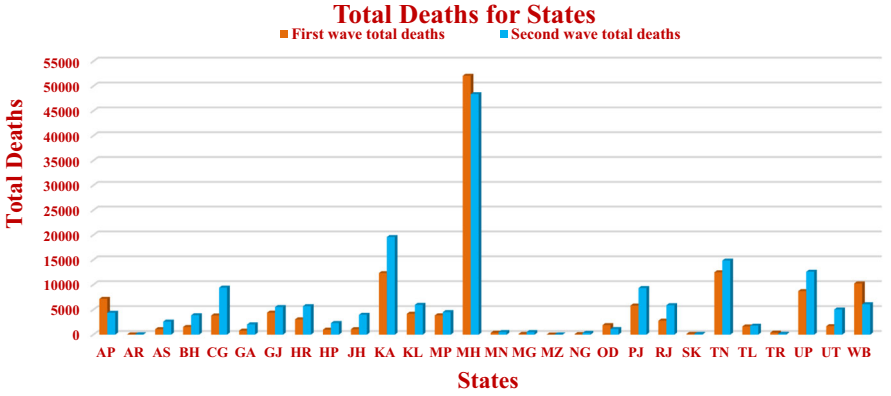

Fig. 5 COVID-19 deaths in states of India during first and second wave

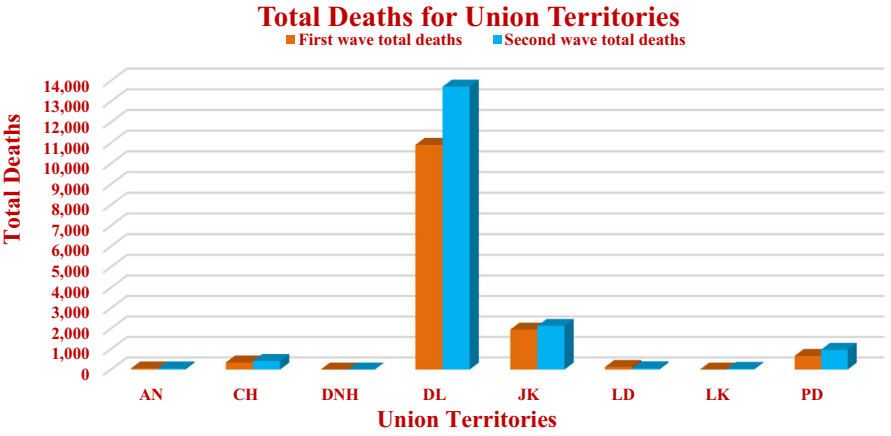

Fig. 6 COVID-19 deaths in union territories of India during first and second wave

January, 2021 in which initially only the frontline workers were given two doses of the vaccine with a gap of 12–16 weeks using two DCGI [1] approved drugs by Oxford-AstraZeneca's Covishield and Bharat Biotech's Covaxin. From 1st March, 2021, the vaccines were given to the residents over the age of 45 and with the onset of the deadly second wave the vaccine is now available for all residents over the age of 18. Table 3 depicts the vaccination status of India as on 8th June, 2021 (Fig. 7).

It is observed from Table 3 that out of the total population of the country, 15.44% people have received the first dose whereas only 3.73% people of the total population has received both the doses. This majorly includes the frontline workers and the residents in the age groups of 45 and more.

Table 3 Vaccination Status of India as on 8th June, 2021

Total residents receiving First Dose	Total residents receiving Second Dose	Age range (years)			Percentage of total population receiving (%)	
		18–45	45–60	60 +	First Dose	Second Dose
18,69,33,771	4,51,29,717	4,72,33,720	7,79,70,082	6,17,29,969	15.44%	3.73%

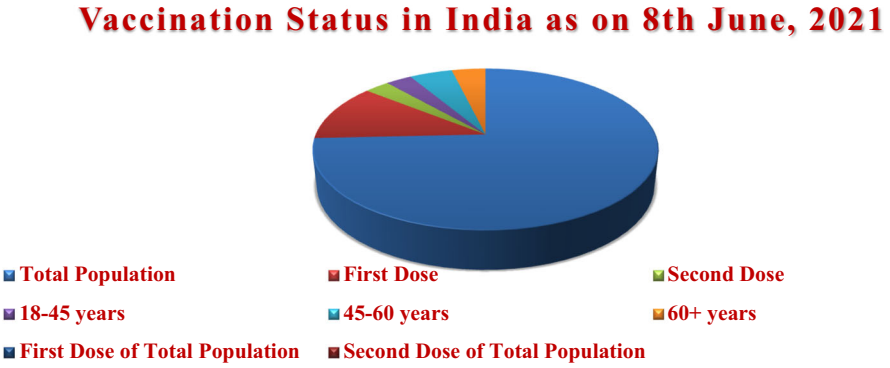

Fig. 7 Vaccination status in India as on 8th June, 2021

3.6 COVID-19 Vaccination Status in States and UTs of India

Table 4 depicts the demographics and the vaccination status of each state and UT of India. The demographic details include the total population of the state/UT, its population density and the rural and urban population percentage (Figs. 8, 9).

The following observations are made from Table 4:

- In states Assam, Bihar, Jharkhand, Meghalaya, Nagaland, Tamil Nadu, Uttar Pradesh and West Bengal out of the total population, less than 15% people have been vaccinated with the first dose
- The states/UTs Dadra and Nagar Haveli, Goa, Himachal Pradesh, Ladakh and Lakshadweep have recorded very good vaccination program having more than 30% vaccinated residents with first dose

4 Cluster Analysis

Cluster analysis is a statistical data mining technique [5] used for grouping the data set having similarities in their parameters. In the present study, two data mining clustering techniques namely, k-means clustering and hierarchical agglomerative clustering

Table 4 Demographics and Vaccination status of Indian states and UTs as on 8th June, 2021

State/UT	Total Population	Density (/km^2)	Rural (%)	Urban (%)	First Dose	Second Dose	Percentage of total population receiving (%)	
							First Dose (%)	Second Dose (%)
AN	3,80,581	46	62	38	1,12,507	3,80,581	29.56	4.01
AP	4,95,77,103	303	71	29	84,29,887	4,95,77,103	17.00	5.21
AR	13,83,727	17	77	23	3,06,022	13,83,727	22.12	5.69
AS	3,12,05,576	398	86	14	35,34,232	3,12,05,576	11.33	2.74
BH	10,40,99,452	1102	89	11	92,77,636	10,40,99,452	8.91	1.78
CH	10,55,450	9252	03	97	3,08,126	10,55,450	29.19	7.13
CG	2,55,45,198	189	77	23	51,55,182	2,55,45,198	20.18	4.45
DNH	5,85,764	970	42	58	2,03,171	5,85,764	34.68	4.45
DL	1,67,87,941	11,297	03	98	44,19,597	1,67,87,941	26.33	7.85
GA	14,58,545	394	38	62	4,85,171	14,58,545	33.26	6.62
GJ	6,04,39,692	308	57	43	1,43,59,541	6,04,39,692	23.76	7.16
HR	2,53,51,462	573	65	35	52,76,368	2,53,51,462	20.81	4.05
HP	68,64,602	123	90	10	21,02,102	68,64,602	30.62	6.33
JK	1,22,67,032	297	74	26	29,32,304	1,22,67,032	23.90	4.64
JH	3,29,88,134	414	76	24	38,64,177	3,29,88,134	11.71	2.27
KA	6,10,95,297	319	61	39	1,26,12,607	6,10,95,297	20.64	4.73
KE	3,34,06,061	859	52	48	83,02,066	3,34,06,061	24.85	6.50

Table 4 (continued)

State/UT	Total Population	Density (/km^2)	Rural (%)	Urban (%)	First Dose	Second Dose	Percentage of total population receiving (%)	
							First Dose (%)	Second Dose (%)
LD	2,74,000	3	16	84	1,42,943	2,74,000	52.17	13.62
LK	64,473	2013	22	78	33,318	64,473	51.68	10.83
MP	7,26,26,809	236	72	28	1,12,99,196	7,26,26,809	15.56	2.55
MH	11,23,74,333	365	55	45	1,95,97,498	11,23,74,333	17.44	4.32
MN	25,70,390	122	70	30	4,33,599	25,70,390	16.87	2.76
MG	29,66,889	132	80	20	4,03,761	29,66,889	13.61	2.51
MZ	10,97,206	52	48	52	2,76,319	10,97,206	25.18	4.69
NG	19,78,502	119	71	29	2,39,568	19,78,502	12.11	2.63
OD	4,19,74,219	269	83	17	69,38,186	4,19,74,219	16.53	3.63
PD	12,47,953	2598	32	68	2,42,580	12,47,953	19.44	4.16
PJ	2,77,43,338	551	63	37	41,63,128	2,77,43,338	15.01	2.88
RJ	6,85,48,437	201	75	25	1,46,15,154	6,85,48,437	21.32	4.75
SK	6,10,577	86	75	25	1,80,310	6,10,577	29.53	9.78
TN	7,21,47,030	555	52	48	80,31,685	7,21,47,030	11.13	2.92
TL	3,50,03,674	312	61	39	55,54,284	3,50,03,674	15.87	3.75
TR	36,73,917	350	74	26	11,32,397	36,73,917	30.82	13.83
UP	19,98,12,341	828	78	22	1,70,79,924	19,98,12,341	8.55	1.83
UT	1,00,86,292	189	70	30	23,96,288	1,00,86,292	23.76	6.83
WB	9,12,76,115	1029	68	32	1,24,92,937	9,12,76,115	13.69	4.34

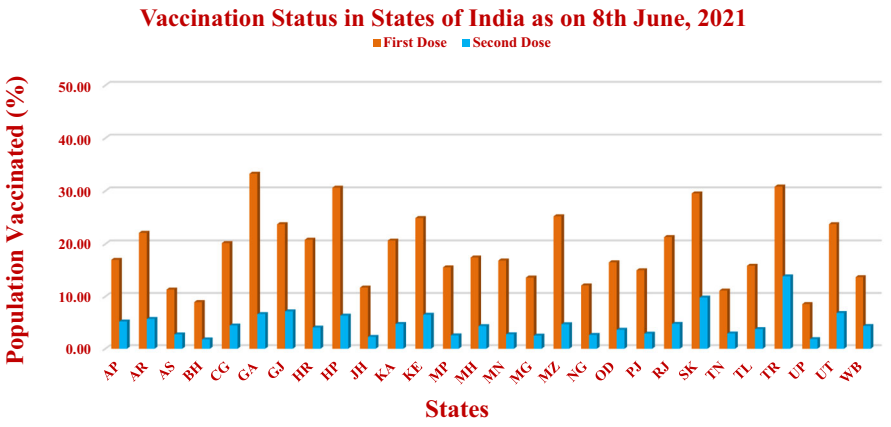

Fig. 8 Vaccination status in states of India as on 8th June, 2021

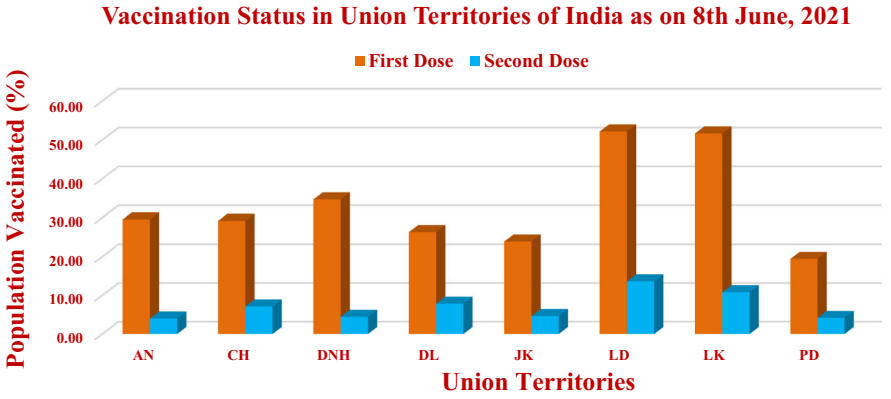

Fig. 9 Vaccination status in union territories of India as on 8th June, 2021

is applied on the COVID-19 dataset for grouping the states/UTs based on their demographics, number of COVID confirmed and death cases and the vaccination status.

4.1 *k-means* Clustering

k-means clustering [5] is an unsupervised clustering technique in which the dataset is partitioned into *k* clusters such that the variance between the dataset within the cluster is minimum. Each data from the dataset belongs to a cluster with the nearest mean. The *k-means* algorithm returns the average value of the parameters. In the present study, *k-means* clustering is applied to group the states/UTs based on three cases:

- *Case A_kclust* Clustering the states/UTs based on the total number of COVID-19 cases and deaths during the first and the second wave

- *Case B_kclust* Clustering of states/UTs to observe the vaccination status with respect to their population
- *Case C_kclust* Clustering the states/UTs to group them respective to the number of COVID-19 cases and deaths with their vaccination status

The following steps are performed while implementing the *k-means* clustering algorithm:

1. Determination of the parameters used for clustering the data set into groups
2. Implementation of elbow method [2, 5] for determining the optimal number of clusters. The elbow method also known as knee of curve method is a heuristic approach used to determine the number of clusters in a dataset.
3. Applying *k-means* clustering using the optimal number of clusters determined in the elbow method.

Case A_kclust: Clustering the states/UTs based on the number of COVID-19 cases and deaths.

Four parameters considered while performing the clustering operation are: total confirmed cases during first wave, total deaths during first wave, total confirmed cases during second wave and total deaths during second wave. Figure 10 depicts the result of the elbow method applied on the dataset.

It is observed from Fig. 10 that the total within the sum of squares value does not vary much after 3 clusters and so 3 is considered as the optimal number of clusters. By applying k-means algorithm with k = 3 gives cluster of sizes 28, 7 and 1. The result and the plotting of k-means clustering is tabulated in Table 5 and Fig. 11 respectively.

The states belonging to each cluster are depicted in Table 6:

The following inferences are made from the result of k-means clustering depicted in Table 5 and 6:

- Maharashtra is the worst hit state with maximum number of COVID-19 confirmed cases and deaths during both the waves
- Although there is a substantial increase in the number of COVID-19 cases in the states belonging to cluster 2, overall, they are moderately affected
- The states belonging to cluster 1 have fewer number of cases as compared to those in cluster 2

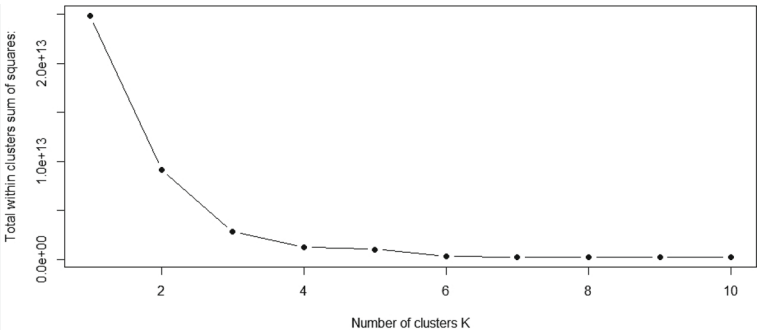

Fig. 10 Elbow method plot for obtaining optimal number of clusters for Case A_kclust

Table 5 Result of k-means clustering algorithm for Case A_kclust

Cluster	First wave total cases	First wave total deaths	Second wave total cases	Second wave total deaths
1	1,20,884	1388	2,08,681	2452
2	7,95,026	9438	11,95,022	11,033
3	21,46,777	52,092	36,95,223	48,378

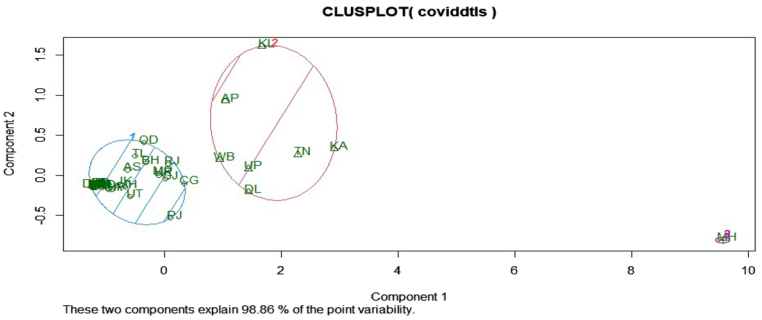

Fig. 11 k-means cluster plot for Case A_kclust

Table 6 States/ UTs belonging to Cluster 1, 2 and 3 for Case A_kclust

Cluster 1	Cluster 2	Cluster 3
AN, AR, AS, BH, CH, CG, DNH, GA, GJ, HR, HP, JK, JH, LD, LK, MP, MN, MG, MZ, NG, OD, PD, PJ, RJ, SK, TL, TR, UT	AP, DL, KA, KL, TN, UP, WB	MH

Case B_kclust Clustering the states/UTs based on their vaccination status.

The parameters considered while performing the clustering operation is the percentage of residents vaccinated by the first and the second dose out of the total population. The elbow method is used to determine the optimal number of clusters as depicted in Fig. 12.

The elbow method suggests 3 as the optimal number of clusters. k-means algorithm applied on the data set with k = 3 gives cluster of sizes 18, 4 and 14. Table 7 depicts the mean of the percentage of people vaccinated by first and second dose in each cluster. Figure 13 shows the plotting of k-means clustering for Case B_kclust.

The states belonging to each cluster are depicted in Table 8:

The following observations are noted from the results derived in Tables 7 and 8.

- The states belonging to cluster 1 have more than 50% and 10% vaccinated residents with first and second dose respectively
- The states belonging to cluster 2 are the least vaccinated states

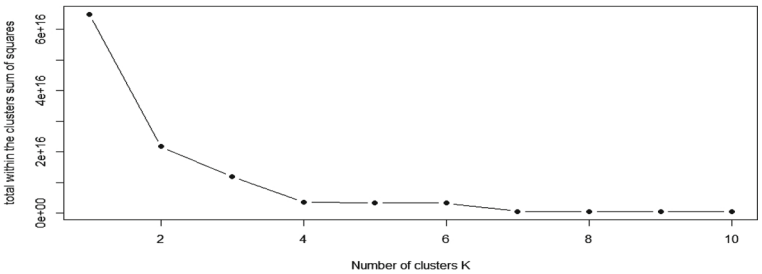

Fig. 12 Elbow method plot for obtaining optimal number of clusters for Case B_kclust

Table 7 Result of k-means clustering algorithm for Case B_kclust

Cluster	First dose (%)	Second dose (%)
1	51.93	12.23
2	15.39	3.41
3	27.68	6.82

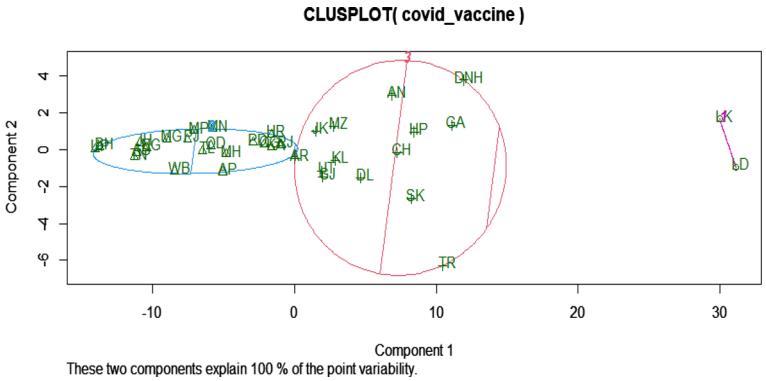

Fig. 13 k-means cluster plot for Case B_kclust

Table 8 States/ UTs belonging to Cluster 1, 2 and 3 for Case B_kclust

Cluster 1	Cluster 2	Cluster 3
LD, LK	AP, AS, BH, CG, HR, JH, KA, MP, MH, MN, MG, NG, OD, PD, PJ, RJ, TN, TL, UP, WB	AN, AR, CH, DNH, DL, GA, GJ, HP, JK, KL, MZ, SK, TR, UT

- The first dose vaccination rate of the states/UTs in this cluster is less than 20% which is quiet alarming

- The average vaccination rate of the states/UTs belonging to cluster 3 is moderate with 27% of first dose and 6% of second dose

Case C_kclust Clustering the states/UTs based on their COVID-19 cases and deaths with their vaccination status.

The parameters considered while applying k-means clustering are: total confirmed cases and deaths during the first and the second wave and percentage of residents receiving first and second dose of the vaccine. Similar to Case A_kclust and Case B_kclust, the elbow method shows 3 as the optimal number of clusters as depicted in Fig. 14. Applying k-means algorithm with k = 3, three cluster of sizes 7, 1and 28 are formed. The result and the cluster plot for Case C_kclust is depicted in Table 9 and Fig. 15.

The states belonging to each cluster are depicted in Table 10:

The following observations are made considering Tables 9 and 10.

- Maharashtra the only state belonging to cluster 2, has highest number of COVID-19 cases and deaths in both waves with quiet a smaller number of residents getting vaccinated
- The states/UTs in cluster 2 are moderately affected with COVID-19 cases and deaths having less vaccination rate
- The states/UTs belonging to cluster 3 have least number of COVID-19 cases as compared to the other two clusters with highest vaccination rate

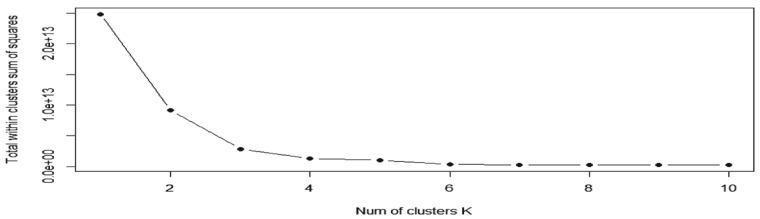

Fig. 14 Elbow method plot for obtaining optimal number of clusters for Case C_kclust

Table 9 Result of k-means clustering algorithm for Case C_kclust

Cluster	First wave total cases	First wave total deaths	Second wave total cases	Second wave total deaths	First dose (%)	Second dose (%)
1	7,95,026	9438	11,95,022	11,033	17.45	4.76
2	21,46,777	52,092	36,95,223	48,378	17.44	4.32
3	1,20,884	1388	2,08,681	2452	23.55	5.37

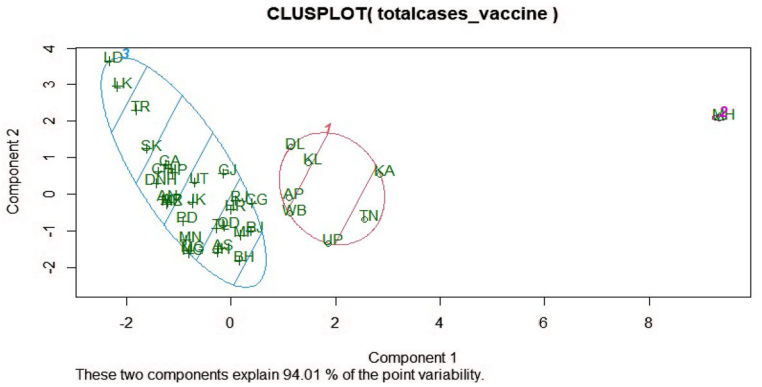

Fig. 15 k-means cluster plot for Case C_kclust

Table 10 States/ UTs belonging to Cluster 1, 2 and 3 for Case C_kclust

Cluster 1	Cluster 2	Cluster 3
AP, DL, KA, KL, TN, UP, WB	MH	AN, AR, AS, BH, CH, CG, DNH, GA, GJ, HR, HP, JK, JH, LD, LK, MP, MN, MG, MZ, NG, OD, PD, PJ, RJ, SK, TL, TR, UT

4.2 Hierarchical Agglomerative Clustering

An unsupervised data mining statistical approach [5] used for grouping data set with similar characteristics by building a hierarchy of clusters. Two types of hierarchical clustering can be performed; i.e., agglomerative and divisive. The type of hierarchical clustering performed in the present study is agglomerative which creates groups from bottom to top. In this method, each observation of the dataset is considered as a cluster which merges with other clusters while moving up the hierarchy. The squared Euclidean distance is used to find the similarities in the data set and average method is used to evaluate the distance between the clusters.

Given two set of points *p (p1, p2)* and *q (q1, q2)*, the Euclidean distance between points p and q is measured as:

$$D(p,q) = \left[(q1 - p1)^2 + (q2 - p2)^2\right]^{1/2}$$

The algorithm is applied to group the states/UTs based on the following three cases:

- *Case A_hclust* Clustering the states/UTs hierarchically based on the total number of COVID-19 cases and deaths during the first and the second wave
- *Case B_hclust* Clustering of states/UTs to observe the vaccination status with respect to their population

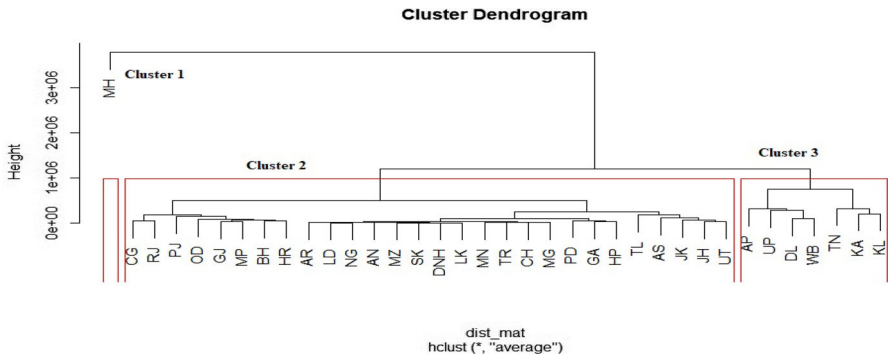

Fig. 16 Dendrograms showing clustering of states/UTs for Case A_hclust

- *Case C_hclust* Clustering the states/UTs to group them respective to the number of COVID-19 cases and deaths with their vaccination status

The following steps are performed while implementing the hierarchical agglomerative clustering algorithm:

1. Determination of the parameters used for clustering the data set into groups
2. Implementation of elbow method [26] for determining the optimal number of clusters.
3. Application of hierarchical agglomerative using the optimal number of clusters determined in the elbow method.
4. Creating the cluster dendrograms and highlighting individual clusters
5. Creating the phylogenetic tree representation for better understanding of the clusters

Case A_hclust Hierarchical clustering is applied with parameters total confirmed cases and deaths during the first and the second wave. The dendrogram showing the clustering of states/UTs for Case A_hclust is depicted in Fig. 16.

Each cluster is marked by the red border. The following observations are made from Fig. 16:

- Maharashtra is the only state belonging to cluster 1 with more than twenty lakh cases in both the waves.
- Twenty-eight states/UTs group together for cluster 2.
- Cluster 2 is formed by merging three sub-clusters.:
  - The first sub-cluster contains eight states having less than 4 lakh cases during the first wave and less than 7 lakh cases during the second wave. The total cases considering both the waves is less than 10 lakhs
  - The second sub-cluster contains 15 states having less than 90,000 cases during the first wave and less than 2 lakh cases during the second wave
  - The third sub-cluster has 5 states having total cases less than 3 lakh during first wave but less than 6 lakh total cases considering both the waves
- Cluster 3 contains 7 states having more than 5 lakh confirmed cases and more than 4000 deaths during both the waves

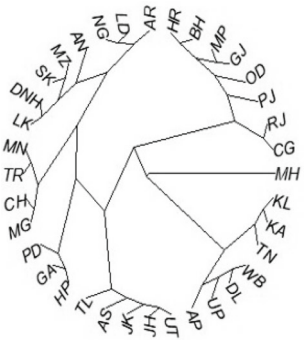

Fig. 17 A phylogenetic tree showing clustering of states/UTs for Case A_hclust

A phylogenetic tree for Case A_hclust using "Radial" representation is depicted in Fig. 17.

Case B_hclust The clustering of states/UTs to observe the vaccination status is done using the percentage of residents receiving the first and the second dose of the vaccine out of their total population. The dendrogram showing the results is depicted in Fig. 18.

The following observations are made from the dendrogram representing the vaccination status of the depicted in Fig. 18:

- Union territories Ladakh and Lakshadweep belonging to cluster 1 has more than 50% first dose vaccinated residents
- The 15 states/UTs grouped into 7 sub-clusters merge to form cluster 2
  - The first dose vaccination percentage in cluster 2 is less than 20% of their total population
  - The states/UTs in this cluster represent least number of vaccinated residents
- Cluster 3 contains 19 states/UTs grouped by sub-clusters having vaccination percentage between 20 and 50%

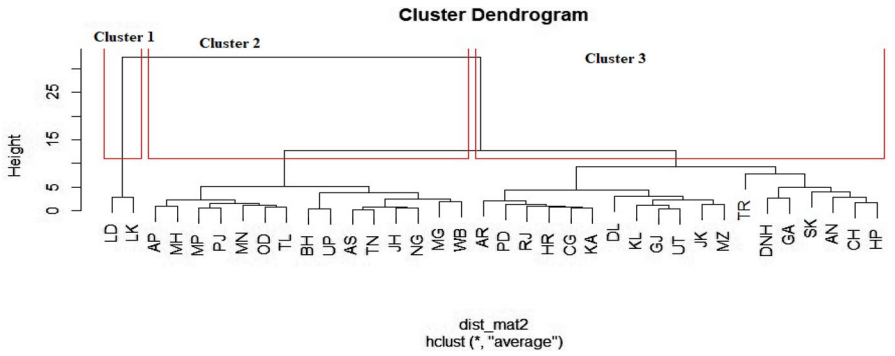

Fig. 18 Dendrogram showing clustering of states/UTs for Case B_hclust

A radial representation of the phylogenetic tree for Case B_hclust is depicted in Fig. 19.

Case C_hclust The clustering of the states/UTs based on their COVID-19 cases and deaths with their vaccination status has the following parameters: total confirmed cases and deaths during the first and the second wave and the percentage of residents receiving first and second dose of the vaccine out of the total population. The dendrogram and radial representation of the phylogenetic tree showing clustering of states/UTs for Case C_hclust is depicted in Figs. 20 and 21 respectively.

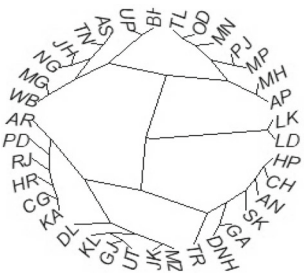

Fig. 19 A phylogenetic tree showing clustering of states/UTs for Case B_hclust

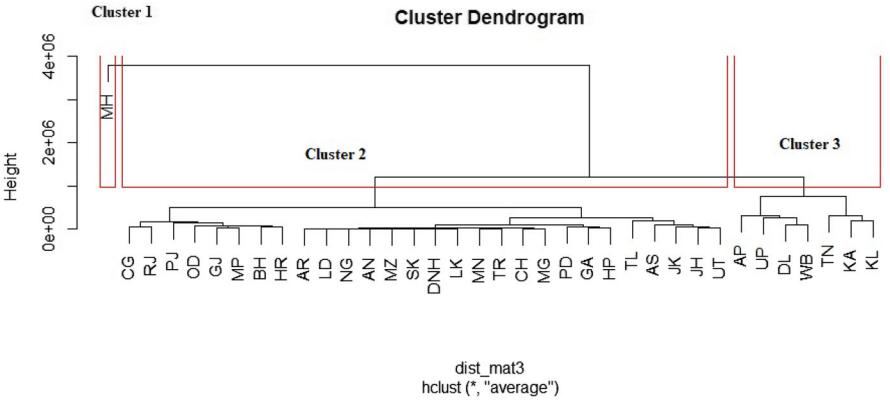

Fig. 20 Dendrogram showing clustering of states/UTs for Case C_hclust

Fig. 21 A phylogenetic tree showing clustering of states/UTs for Case C_hclust

The following observations are made from the dendrogram depicted in Fig. 20:

- Maharashtra is the only state belonging to cluster 1 with highest number of covid-19 cases during both the waves and less than 20% residents getting vaccinated with first dose
- 28 states/UTs grouped into sub-clusters merge to form cluster 2. The sub-clusters are formed by the total number of cases during both waves and their vaccination percentage.
  - First 2 states in the sub-cluster have total cases between 3 and 7 lakhs with more than 20% vaccination
  - Next 6 states/UTs have covid cases up to 3 lakhs during the first wave and more than 3 lakh cases during the second wave with more than 8% vaccination
  - Sub-cluster 3 formed by grouping 15 states/UTs have less than 60,000 cases during the first wave and up to 1.5 lakh cases during the second wave with more than 10% vaccination
  - Sub-cluster 4 comprises of 5 states with cases between 90,000 and 3 lakhs during both the waves and more than 10% vaccination
- The 7 states /UT belonging to cluster 3 have more than 6 lakh COVID-19 cases during both the waves with less than 30% vaccination.

5 Research Findings and Future Scope

The present study performs cluster analysis on the COVID-19 dataset of each state/UT of India from March, 2020 to 8th June, 2021. Two unsupervised clustering algorithms are applied to the COVID-19 and vaccination data set in order to group and monitor the states and the UTs based on their increase/decrease of covid cases, deaths and vaccination status. The situation in the states having maximum number of COVID-19 cases and deaths both during the first and the second wave with less than 20% vaccination is quiet alarming. Worst affected states during both the waves are Maharashtra, Andhra Pradesh, Delhi, Karnataka, Kerala, Tamil Nadu, Uttar Pradesh and West Bengal. Looking into the severity of the second wave, faster vaccination process in the densely populated states may help in reducing the rapid transmission of the infection. The vaccination drive in the rural part of the country is a major challenge as myths regarding the vaccines proves to be a major obstacle.

The results of the analysis presented in this study will provide useful information regarding the pandemic and to the frontline workers combatting the spread of the infection in the country. The results can be used for pursuing further research for the betterment of government policies. The analysis can be further carried out at district level of each state/UT to explore and identify the useful information about the disease spread and the vaccination drive.

Author's contributions The only author.

Funding Not Applicable.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest Not applicable.

Ethical Statements I hereby declare that this manuscript is the result of my independent creation under the reviewers' comments. Except for the quoted contents, this manuscript does not contain any research achievements that have been published or written by other individuals or groups. I am the only author of this manuscript. The legal responsibility of this statement shall be borne by me.

References

1. Ministry of Health and Family Welfare, Government of India (2021). https://www.mohfw.gov.in. Accessed 20 June 2021
2. PRS Legislative Research (2021). https://prsindia.org/covid-19/cases. Accessed 20 June 2021
3. The Office of the Registrar General & Census Commissioner, India (2011). https://censusindia.gov.in. Accessed 20 June 2021
4. World Health Organization (2021). https://www.who.int/emergencies/diseases/novel-coronavirus. Accessed 20 June 2021
5. Han J, Kamber M, Pei J (2011) Data mining-concepts and techniques, 3rd edn. Morgan Kaufman Publishers
6. Olson DL, Shi Y (2007) Introduction to business data mining. McGraw-Hill/Irwin, New York
7. Shi Y, Tian YJ, Kou G, Peng Y, Li JP (2011) Optimization based data mining: theory and applications. Springer, Berlin
8. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4(2):149–178
9. COVID-19 India (2021). https://www.covid19india.org. Accessed 20 June 2021
10. Wang Y, Grunewald M, Pearlman S (2020) Corona viruses: an updated overview of their replication and pathogenesis. Methods Mol Biol. https://doi.org/10.1007/978-1-0716-0900-2_1
11. Dehkordi A, Alizadeh M, Derakshan P, Babazadeh P, Jahandideh A (2020) Understanding epidemic data and statistics: a Case Study of COVID-19. J Med Virol. https://doi.org/10.1101/2020.03.15.20036418
12. Deng Y, He F, Li W (2020) Coronavirus disease 2019: What we know? J Med Virol. https://doi.org/10.1002/jmv.25766
13. Wang J, Tang K, Feng K, Lv W (2020) High temperature and high humidity reduce the transmission of covid-19. https://arxiv.org/abs/2003.05003. Accessed 25 June 2021
14. Gupta S (2020) Epidemic parameters for COVID-19 in several regions of India. https://arxiv.org/abs/2004.11677 Accessed 24 June 2021
15. Gupta S (2020) The age and sex distribution of COVID-19 cases and fatalities in India. https://www.medrxiv.org/content/https://doi.org/10.1101/2020.07.14.20153957. Accessed 24 June 2021
16. Gupta S (2020) Inferring epidemic parameters for COVID-19 from fatality counts in Mumbai. https://arxiv.org/abs/2004.11677. Accessed 24 June 2021
17. Gupta S, Shankar R (2020) Estimating the number of COVID-19 infections in Indian hot-spots using fatality data. https://arxiv.org/abs/2004.04025. Accessed 24 June 2021
18. Kumar S (2020) Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. Ann Data Sci 7:417–425

19. Li J, Guo K, Herrera Viedma E, Lee H, Liu J, Zhong Z, Gomes L, Filip FG, Fang SC, Özdemir MS, Liu XH, Lu G, Sh Y (2020) Culture vs policy: more global collaboration to effectively combat COVID-19. Innov. https://doi.org/10.1016/j.xinn.2020.100023
20. Liu Y, Gu Z, Xia S, Shi B, Zhou X, Shi Y, Liu J (2020) What are the underlying transmission patterns of COVID-19 outbreak? An age-specific social contact characterization. EClincialMedicine 22:100354
21. Temesgen A, Gurmesa A, Getchew Y (2018) Joint modeling of longitudinal CD4 count and time-to-death of HIV/TB co-infected patients: a case of jimma university specialized hospital. Ann Data Sci 5:659–678
22. Hussain A, Bouachir O, Turjman F, Alooqaily M (2020) AI techniques for COVID-19. IEEE Access 8:128776–128795
23. Gondauri D, Mikautadze E, Batiashvili M (2020) Research on covid-19 virus spreading statistics based on the examples of the cases from different countries. Electron J Gen Med 17:em209
24. Kumar J, Agiwal V, Yau C (2021) Study of the trend pattern of COVID-19 using spline-based time series model: a Bayesian paradigm. Jpn J Stat Data Sci: 1–15
25. Pai C, Bhaskar A, Rawoot V (2020) Investigating the dynamics of COVID-19 pandemic in India under lockdown. Chaos Solitons Fractals. https://doi.org/10.1016/j.chaos.2020.109988. Accessed20June2021
26. National Program on Technology Enhanced Learning (NPTEL) (2021). nptel.ac.in/courses/106/106/106106179. Accessed 10 June 2021