RESEARCH ARTICLE

Evaluation of First and Second Dose of COVID-19 Vaccination using k-means Clustering Model and Visualization of Indian States and Union Territories

R Lakshmi Priya[1], Manjula Devi N[2], Manimannan Ganesan[3]

Abstract

Aim: This research paper was attempted to identify the pattern of COVID-19 vaccinated people of first and second doses using machine learning (ML) methods.

Settings and designs: The secondary source of COVID-19 vaccination data was collected from the National Informatics Centre (NIC), India, up to April 30, 2021, based on Census 2011 data. The original data consist of total population, first dose, second dose, percentage of the first dose, percentage of the second dose, and the cumulative percentage of the population throughout the states and union territories of India.

Materials and methods: Application of Orange data mining software determines the clusters and plots of the graph of vaccination data for various states and union territories of India. The file widget opens a new vaccination data set and performs k-means++ from two to nine with silhouette distances.

Results: Silhouette distance scores and cluster information are achieved. The three zones are visualized and the zones are labeled as green, blue, and red. Cluster 1 (C1) zone indicates that states and union territories are highly vaccinated, cluster 2 (C2) zone indicates that states and union territories are moderately vaccinated, and the cluster 3 (C3) zone is low-vaccinated states and union territories of India. The different colors green, blue, and red of the zones are labeled as C1, C2, and C3, respectively.

Conclusion: In India, Sikkim, Tripura, Ladakh, and Lakshadweep have a low population density but fall under the highly vaccinated zones of first and second doses. Goa, Mizoram, Delhi, Arunachal Pradesh, Chandigarh, Uttarakhand, Gujarat, Rajasthan, Kerala, Jammu and Kashmir, Dadra and Nagar Haveli, Damn and Diu, Himachal Pradesh, Chhattisgarh, and Andaman Nicobar Islands have diverse population density and come in the category of low-vaccinated zones of first and second doses. Manipur, Meghalaya, Nagaland, Odisha, West Bengal, Haryana, Karnataka, Andhra Pradesh, Maharashtra, Telangana, Jharkhand, Madhya Pradesh, Punjab, Assam, Uttar Pradesh, Tamil Nadu, Puducherry, and Bihar have high population density and are considered under moderately vaccinated zones of first- and second-dose COVID-19 vaccination.

Keywords: COVID-19 vaccination, Data science and visualization, k-means, NFHS, Silhouette distance score.

Journal of Health Sciences & Research (2021): 10.5005/jp-journals-10042-1104

Introduction

A number of thoughts and recommendations to support the data and the way it can be efficiently utilized in current pandemic situation.[1] Quarantine is no longer a separation of those who don't seem to be sick but believed to have an exposure of infection, to prevent from transmission of illnesses. Human beings are typically isolated from their homes; however they may also be quarantined in community-based motels. Wondering about the growing extent of various victims and restrained community-based facilities, the regular public are being requested to quarantine in their houses.

The cluster containment technique would be used to manage the disease under a described geographic vicinity by way of early vaccination of populations, breaking the chain of transmission, and subsequently stopping its unfoldment to new areas. This will embody the percentage of the first- and second-dose vaccination and ordinary measures. Many touchy factors are associated while defining the containment zones in the states and union territories.

Manually defining, these altering clusters in dimension and vicinity may not always be viable. The strategies from data science, particularly Ok approach, can assist in this scenario via defining the clusters as vaccinated states and union territories. The intention of this paper was to deliver neutrality and accuracy in developing the COVID-19 vaccination zones dynamically and constantly utilizing the Ok method of information technological know-how that is suited to get a nonstop update on correct and contemporary microdegree isolation data of vaccination zones to devise the impartial strategies for putting aside contacts of COVID-19 victims from network. The scope of this research paper was to practice k-means++ method from information technological know-how on the gathered vaccination facts like first and second doses of vaccination, complete populace to outline, and visual plot of the vaccinated zones on the true map.

[1]Department of Statistics, Dr. Ambedkar Government Arts College, Chennai, Tamil Nadu, India

[2]Department of Community Medicine, Karpaga Vinayagar Medical College, Chengalpattu, Tamil Nadu, India

[3]Department of Mathematics, TMG College of Arts and Science, Kanchipuram, Tamil Nadu, India

Corresponding Author: Manimannan Ganesan, Department of Mathematics, TMG College of Arts and Science, Kanchipuram, Tamil Nadu, India, e-mail: manimannang@gmail.com

How to cite this article: Priya RL, Manjula Devi N, Ganesan M. Evaluation of First and Second Dose of COVID-19 Vaccination using k-means Clustering Model and Visualization of Indian States and Union Territories. J Health Sci Res 2021;x(x):xx–xx.

Source of support: Nil

Conflict of interest: None

Background of The Study

All through the COVID-19 disaster, the discipline of data science is in the epicenter.[2] Majority of the public are interested in gazing and searching ahead of the statistical evaluation and epidemiology graphs and sharing equally in social media on a massive scale. The chance from data science is very high. Data science is a growing subject that consists of a quantity of fantastic and beneficial tools, features, and techniques.

The study cautioned the cluster containment approach for Zika virus outbreak was once located high quality in Rajasthan, India. It is defined that how surveillance techniques are used to manipulate the ailment from spreading past containment zones of a 3 km radius.[3] This article offers an emphasis on growing containment to stop the outburst of disease; then again, it does no longer provide an explanation for how to make these zones shortly and accurately.

Additionally, it is defined about the fantastic containment to manage COVID-19 instances in China specifically.[4] The mannequin which they defined in their paper captures each isolation of symptomatic contaminated humans and different populace isolation practices. The center of attention of the lookup function is considered on the contagion method and universal consequences as nicely as the importance of the containment. Their lookup work implies and helps to outline the containment zones accurately.

This research article predicts and classifies the facts of COVID-19 primarily based on the four desktop mastering algorithms with four principal parameters such as validated cases, recoveries, deaths, and energetic cases. The secondary sources of the database had been accumulated from the Ministry of Health and Family Welfare Department (MHFWD) from the states and union territories of India up to March 2021. Based on this background, the database is categorized and anticipated more than a few computing devices gaining knowledge from the algorithms, such as support vector machine (SVM), k-nearest neighbor (kNN), random forest, and logistic regression. Initially, k-means clustering evaluation is used to operate and recognize five significant clusters and is labeled as Very Low, Low, Moderate, High, and Very High of five foremost parameters based totally on their common values.[5]

Additionally, the five clusters validated the usage of four laptop algorithms and the affected states are visualized on the desk with the assistance of prediction and probabilities. The specific computing software study of validation and classification accuracy are 88%, 97%, 91% and 91%. The classification of the states and union territories has been named as Very Low Affected (VLA), Low Affected (LA), Moderately Affected (MA), Highly Affected (HA), and Very Highly Affected (VHA) via COVID-19 cases. Maharashtra is effectively categorized as VHA states; Delhi, Uttar Pradesh, and West Bengal fall under MA category; Assam, Bihar, Chhattisgarh, Haryana, Gujarat, Madhya Pradesh, Odisha, Punjab, Rajasthan, and Telangana fall under the LA States; and Tamil Nadu, Kerala, Andhra Pradesh, and Karnataka types a team of Highly Affected States. The remaining states and union territories fall under VLA by way of COVID-19 cases.

Application of Data Science

Data science generally has a five-stage life cycle and the stages are given in Fig. 1.

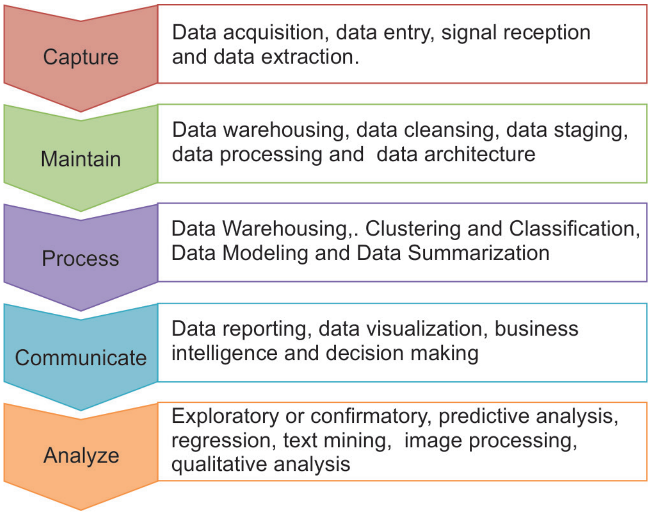
Capture
Data acquisition, data entry, signal reception and data extraction.
Maintain
Data warehousing, data cleansing, data staging, data processing and data architecture
Process
Data Warehousing,. Clustering and Classification, Data Modeling and Data Summarization
Communicate
Data reporting, data visualization, business intelligence and decision making
Analyze
Exploratory or confirmatory, predictive analysis, regression, text mining, image processing, qualitative analysis

Fig. 1: Five stages of data science

The research paper explains how the seismic supply zones have been created using the k-means cluster evaluation for Aegean Region.[6] The paper describes the magnitude of making use of the k-means algorithm for hierarchical cluster evaluation and used to be discovered beneficial for partition areas primarily based on the found seismicity to have a constant method to supply mannequin development. Clustering of various states and union territories of India is prepared which gives the silhouette distances between each of the pairs of the cases as shown in Table 1.[7]

The main objective of this research paper was to explore the applications of the k-means technique of data science influence toward defining, visualizing, and maintaining the vaccination states impacted by COVID-19.

Materials and Methods

The secondary source of vaccination data was collected from NIC, India, up to April 30, 2021, based on Census 2011 population data. The original data consist of total population, first dose, second dose, percentage of the first dose, percentage of the second dose, and the cumulative percentage of the population throughout the states and union territories of India. This research paper used only three parameters such as total population, percentage of the first dose, and percentage of the second dose.[8]

In the methodology, the researcher uses python-based Orange data mining software to identify the vaccinated states of India using k-means clustering techniques with the help of silhouette distance matrices.

The widget applies the k-means clustering algorithm to the data and gives outputs as a new data set in which the clustered index is used as a class attribute. The original class attribute, if it exists, is moved to meta-attributes. Silhouette distance scores of clustering results for various k are also shown in the widget (Table 2).

Machine Learning Algorithms

The proposed Orange data mining software records mining algorithm to execute k-means clustering steps as follows:

Step 1: Initially, select the quantity of clusters with constant algorithm of cluster information in the precise range of clusters.

Step 2: Select the initialization technique of k-means++. The first core is selected randomly; the subsequent are chosen from the closing factors with a chance proportioned to squared distance from the closest center.

Step 3: Run the algorithm to most of the generation with each cluster, and set it manually.

Table 1: Cluster information using k-means++ with silhouette distances

	Pop Census 2011	*Stateunionterntion*	*Cluster*	*Silhouette*	*First_percentage*	*Second_percentage*
1	390581.0	Andaman and...	C1	0.638057	17.7495	1.8075
2	49577103.0	Andhra Pradesh	C2	0.688111	8.09476	1.20999
3	1383727.0	Arunachal Pradesh	C1	0.629764	9.78495	2.69359
4	31205576.0	Assam	C2	0.683697	4.14573	0.943655
5	104099452.0	Bihar	C2	0.682619	4.78915	0.67443
6	1055450.0	Chandigarh	C1	0.657515	11.3222	2.47288
7	25545198.0	Chhattisgarh	C1	0.661322	17.1465	1.98895
8	343709.0	Dadra and Nag..	C1	0.556202	10.6695	1.92692
9	243247.0	Daman and Diu	C1	0.659671	14.6736	1.92549
10	16787941.0	Delhi	C1	0.675566	12.8801	2.73401
11	1458545.0	Goa	C1	0.628213	13.6004	3.16816
12	60439662.0	Gujarat	C1	0.684097	17.8778	2.41419
13	25351462.0	Haryana	C2	0.62867	10.811	1.2842
14	6864602.0	Himachal Pradesh	C1	0.665122	16.7971	1.98763
15	12267032.0	Jammu and Kashmir	C1	0.584099	11.3036	1.92008
16	32988134.0	Jharkhand	C2	0.695922	7.3577	1.05573
17	61095297.0	Karnataka	C2	0.637188	10.6924	1.25082
18	3346061.0	Kerala	C1	0.672189	15.5355	2.03686
19	274000.0	Ladakh	C3	0.689766	24.5201	3.89489
20	644730.0	Lakshadweep	C3	0.663558	23.4594	4.80666
21	72626809.0	Madhya Pradesh	C2	0.683027	9.04137	1.05119
22	112374333.0	Maharashtra	C2	0.667914	9.75275	1.16965
23	2570390.0	Manipur	C2	0.611044	7.00706	1.95017
24	2966889.0	Meghalaya	C2	0.664248	4.23376	1.48192
25	1097206.0	Mizorem	C1	0.654146	11.1573	2.89499
26	1978502.0	Nagaland	C2	0.663511	5.53803	1.54339
27	41974219.0	Odisha	C2	0.609048	10.272	1.47422
28	1247953.0	Puducherry	C2	0.595827	12.1546	1.11887
29	277443338.0	Punjab	C2	0.686325	7.94818	0.767903
30	68548437.0	Rajasthan	C1	0.668851	13.799	2.07999
31	610577.0	Sikkim	C3	0.674875	22.9486	3.42938
32	72147030.0	Tamil Nadu	C2	0.691738	5.67154	0.859428
33	35003674.0	Telangana	C2	0.685855	7.40264	1.05448
34	3673917.0	Tripura	C3	0.611522	20.9457	3.17378
35	199812341.0	Uttar Pradesh	C2	0.685419	4.55745	0.806766
36	10096292.0	Uttarakhand	C1	0.679639	13.1523	2.39523
37	91276115.0	West Bengal	C2	0.681095	8.3188	1.28302

Table 2: k-means of silhouette scores

Silhouette scores	
2	0.503
3	0.544
4	0.491
5	0.432
6	0.411
7	0.433
8	0.409
9	0.418

Step 4: The output widget gives a new vaccination statistics set with appended cluster statistics and pick how to append cluster facts and identify the column.

Step 5: Tick Apply Automatically to commit the widget adjustments automatically. Otherwise, click Apply.

Step 6: Produce a report.

Step 7: Check the ratings of clustering consequences for more than a few k[9] (Figs 2 to 7).

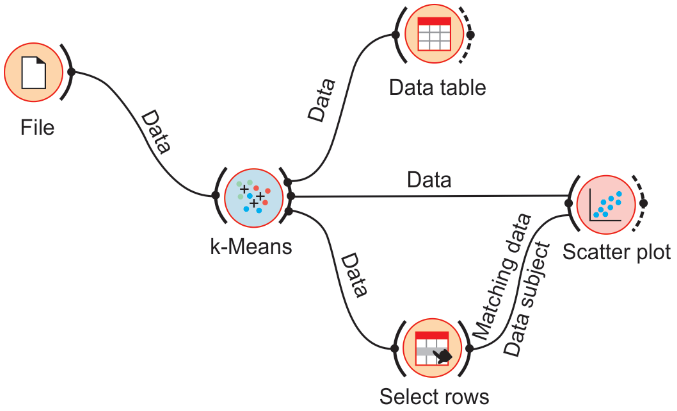

Fig. 2: k-means data mining workflow widget

Results and Discussions

The application of the Orange data mining software program is to decide the clusters and visualize the layout of vaccination records for more than a few states and union territories. The file widget opens new vaccination information set and functions k-means++ from two to nine with silhouette distances. The silhouette rankings are

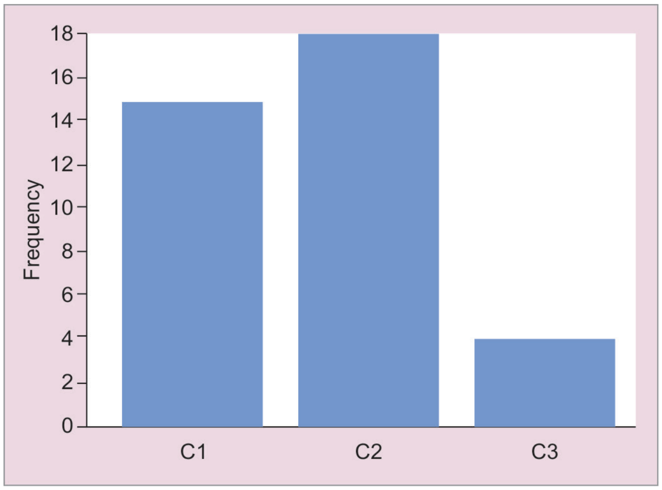

Fig. 4: Data distribution of three clusters of the states and union territories of India (vaccination data)

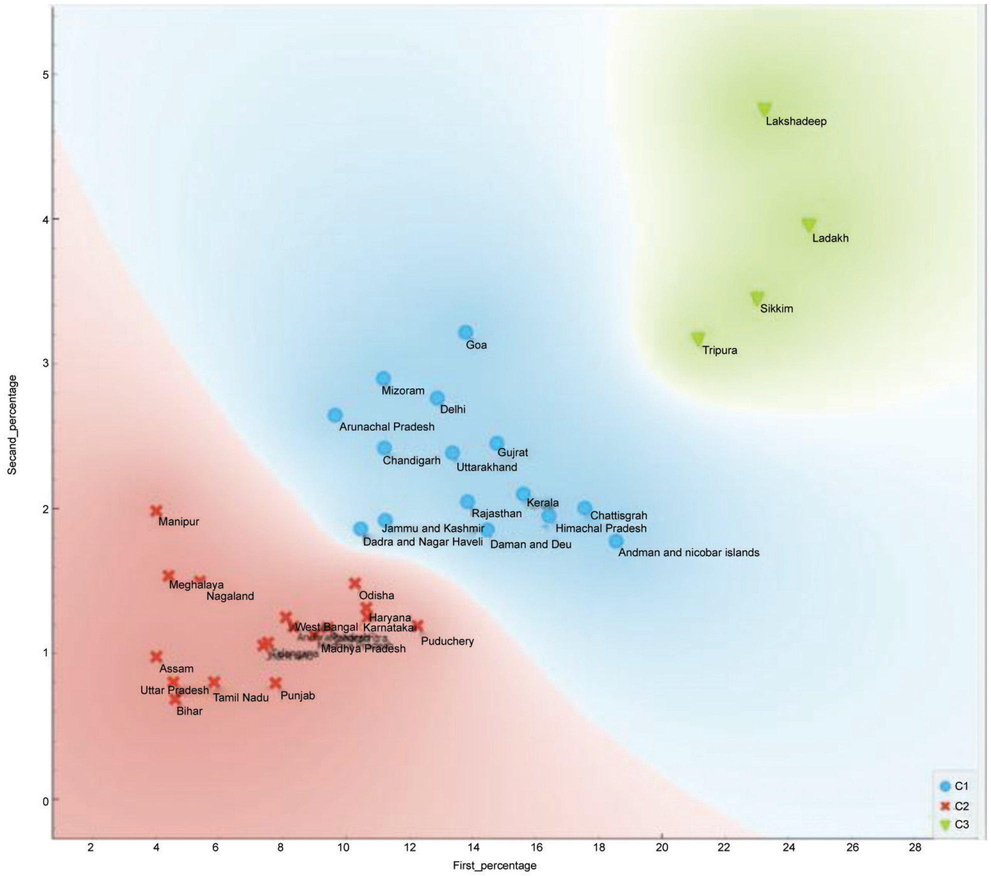

Fig. 3: Data visualization of three clusters of the states and union territories of India (vaccination data)

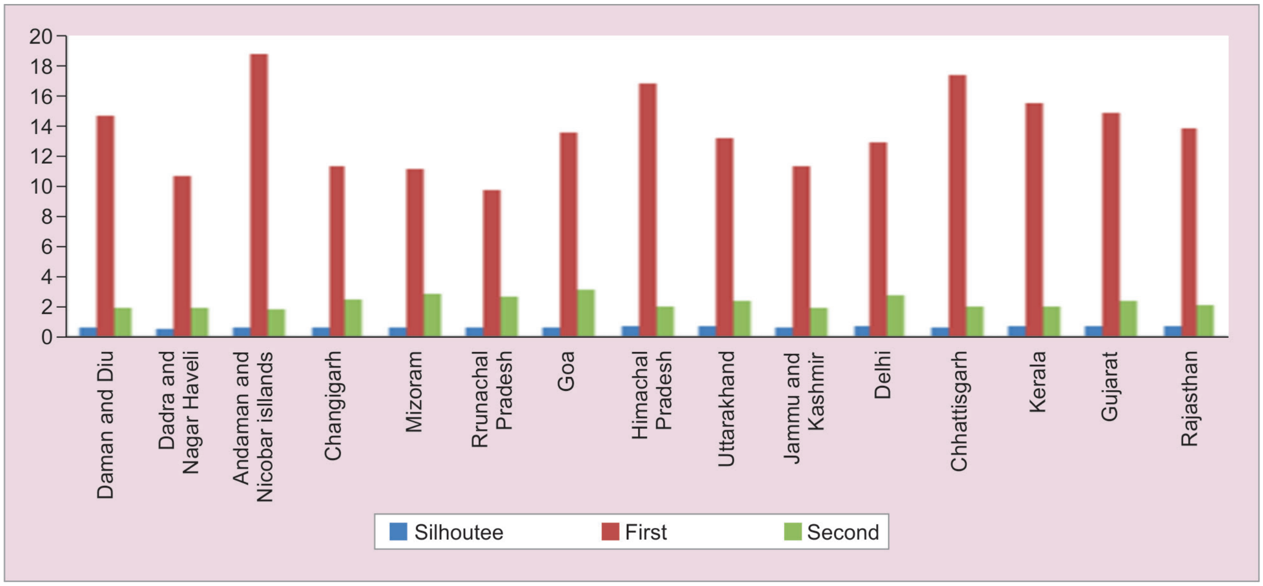

Fig. 5: Data distribution (cluster 1) of the states and union territories of India (vaccination data)

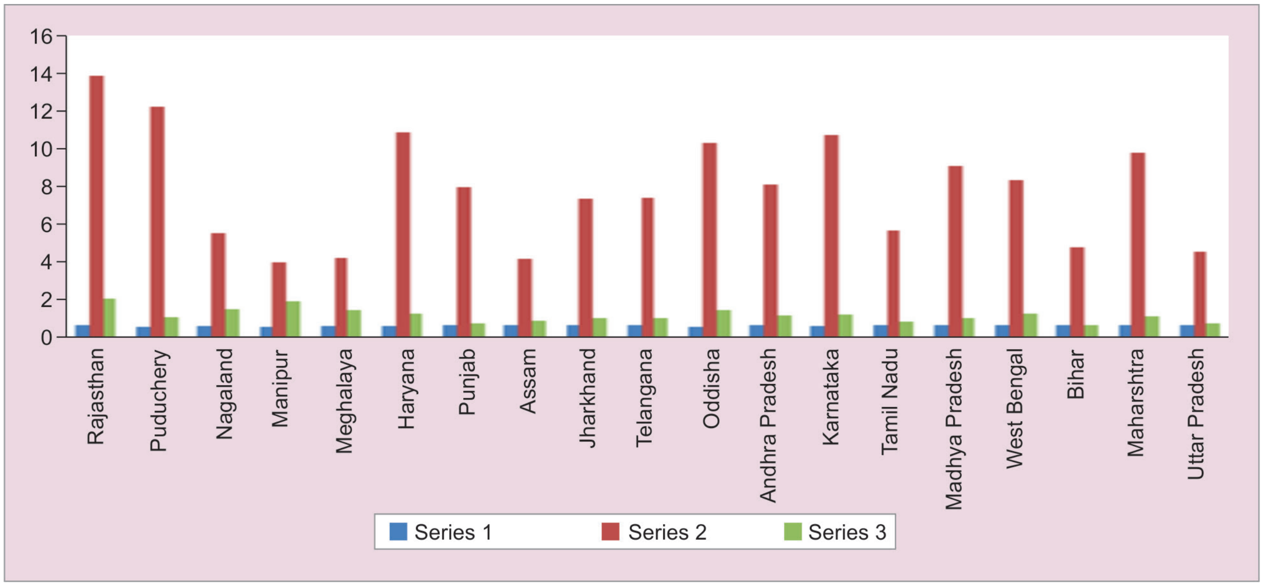

Fig. 6: Data distribution (cluster 2) of the states and union territories of India (vaccination data)

displayed in Table 2. The outcomes of cluster statistics are exhibited in Table 1. The k-means clusters provided three cluster zones of states and union territories primarily based on the first- and second-dose vaccination. The three zones are visualized in the range of three to seven. The inexperienced areas, i.e., green zones, of the states and union territories are highly vaccinated, the blue zones in the states and union territories are moderately vaccinated, and the red zones are mentioned as low-vaccinated states and union territories of India.

Sikkim, Tripura, Ladakh, and Lakshadweep have low populace and fall under highly vaccinated. Goa, Mizoram, Delhi, Arunachal Pradesh, Chandigarh, Uttarakhand, Gujarat, Rajasthan, Kerala, Jammu and Kashmir, Dadra and Nagar Haveli, Daman and Diu, Himachal Pradesh, Chhattisgarh, and Andaman Nicobar Islands have sufficient populace and fall under low-vaccinated zones.

And, Manipur, Meghalaya, Nagaland, Odisha, West Bengal, Haryana, Karnataka, Andhra Pradesh, Maharashtra, Telangana, Jharkhand, Madhya Pradesh, Punjab, Assam, Uttar Pradesh, Puducherry, and Bihar have excessive populace, however fall under rather vaccinated of first and second doses.

Findings and Recommendations

A open-source tool like Orange data mining is found useful for exploring the appropriate and applicable functions in data science. Several partitions with different values of k-means of clusters or partitions are recommended to review along with cluster quality index for an optimum solution. k-means clustering can be improved to a small separation of suppression states and union territories.

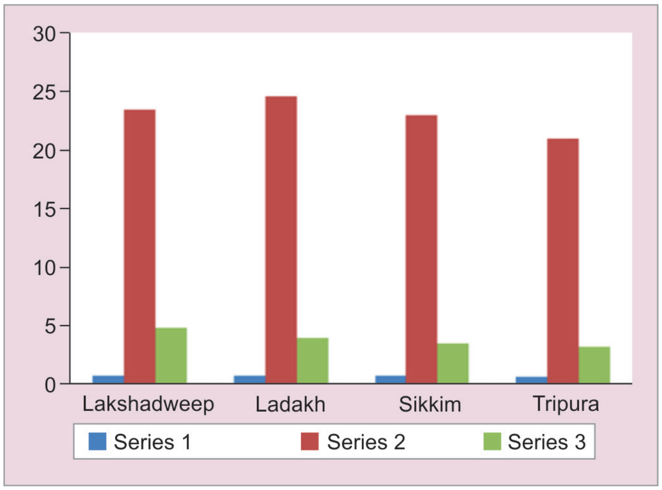

Fig. 7: Data distribution (cluster 3) of the states and union territories of India (vaccination data)

Conclusion

To attain unique results of vaccination assessment using k-means cluster and achieved three meaningful clusters for vaccination data set of the states and union territories of India. It is recommended that data science techniques such as k-means clustering can be adopted to define the microlevel segregation of vaccination zones and manage them effectively. The clusters formed based on COVID-19 patients' vaccination data using data science techniques, specifically k-means++, will be active, accurate, visible, and easy to apply for any given data set.

Orcid

Manimannan Ganesan https://orcid.org/0000-0001-7715-5092

References

1. Callaghan S. COVID-19 is a data science issue. Patterns 2020;1(2):100022. DOI: 10.1016/j.patter.2020.100022.
2. Wollersheim BC. Surprising side effect of COVID-19 : we are all data scientists now. In: Data Analytics & Insights, Arcadis. 2020. Available from: https://ischoolonline.berkeley.edu/data-science/class-profile/.
3. Singh R, Gupta V, Malhotra B, et al. Cluster containment strategy: addressing Zika virus outbreak in Rajasthan, India. BMJ Glob Health 2019;4(5):e001383. DOI: 10.1136/bmjgh-2018-001383.
4. Maier BF, Brockmann D. Effective containment explains sub exponential growth in recent confirmed COVID-19 cases in China. Science 2020;368(6492):742–746. DOI: 10.1126/science.abb4557.
5. Arumugam P, Kadhirveni V, Lakshmi Priya R, et al. Prediction, cross validation and classification in the presence COVID-19 of Indian states and union territories using machine learning algorithms. Int J Recent Technol Eng 2021;10(1):16–20. DOI: 10.35940/ijrte.A5659.0510121.
6. Weatherill G, Burton PW. Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region. Geophys J Int 2008;176(2):565–588. DOI: 10.1111/j.1365-246X.2008.03997.x.
7. Weatherill G, Burton PW. Delineation of shallow seismic source zones using k-means cluster analysis, with application to the Aegean region. Geophys J Int 2009;176(2):565–588. DOI: 10.1111/j.1365-246X.2008.03997.x.
8. National Information Centre (NIC), India. Available from: https://www.mygov.in/covid-19/.
9. Orange Data Mining workflow and Document. Available from: https://orange3.readthedocs.io/en/3.5.0/widgets/unsupervised/kmeansclustering.html.