Clustering Models for COVID-19 Database of Indian States and Union Territories using Machine Learning Methods

Krithika Subbiah[1] N. Manjula Devi[2] R. Lakshmi Priya[3] and Manimannan G.[4]

[1]Post Graduate Student, School of Public Health Sciences, University of Waterloo, Canada.
[2]Biostatistician, Department of Community Medicine, Karpagav Vinayagar Medical College, Chengalpet.
[3]Assistant Professor, Department of Statistics, Dr. Ambedkar Government Arts College, Vyasarpadi, Chennai.
[4]Assistant Professor, Department of Mathematics, TMG College of Arts and Science, Chennai.

Abstract

Background: This research paper attempts to identify the Application of Orange Data mining software that determines hierarchical clusters and plots dendrogram of Total, Positive and Negative sample data for various states and union territories. The Secondary sources of data were collected from April 2020 to April 2021 with the help of three main parameters namely Total Cases, Negative Cases and Positive Cases.

Methodology: Subsequently the python based Orange data mining workflow executed hierarchical clustering methods of Single Linkage, Complete Linkage, Weighted Linkage, Average Linkage and Wards method. The file widget open new COVID-19 data set and perform hierarchical cluster with Euclidean distance measure. The Euclidean distance measure achieved five natural clusters.

Result: The five clusters are visualized in the form of dendrogram and the states and union territories are labeled as five different colors and are labeled as C1, C2, C3, C4 and C5.

Conclusion: The C1 zone indicates that state Uttar Pradesh have Very High (VH) total, positive and negative cases of cluster, C2 zone indicates Bihar, Karnataka, Maharashtra and Tamilnadu states have High (H) total, positive cases and negative cases. C3 zone indicates Andhra Pradesh possess Low (L) total, positive cases and negative cases. C4 zone indicates that the states and union territories of Andaman and Nicobar Islands, Arunachal Pradesh, Chandigarh, Dadra and Nagar Haveli and Daman and Diu, Goa, Himachal Pradesh, Ladakh, Lakshadweep, Manipur, Meghalaya, Mizoram, Nagaland, Puduchery, Sikkim, Tripura and Uttarakhand recorded Very Low (VL) Total, Positive and Negative cases. The final Cluster C5 zone indicates that the states and union territories of Assam, Chhattisgarh, Delhi, Gujarat, Haryana, Jammu and Kashmir, Jharkhand, Kerala, Madhya Pradesh, Odisha, Punjab, Rajasthan, Telengana and West Bengal have Moderate Total, Positive and Negative cases. The open source tools like Orange Data mining found useful for exploring the appropriate and applicable functions in the data science. Several clustering methods are recommended to review along with cluster Euclidean distance for optimum solution. The clusters formed based on COVID-19 patient's Total, Positive and Negative cases data using Data Science techniques specifically. Hierarchical clustering methods will be active, unbiased, accurate, visible, economic and easy to apply.

Keywords: *Total Cases, Negative Cases, Positive Cases, Machine Learning methods, Hierarchical Clustering, Dendrogram, COVID-19 Indian States and Union Territories.*

Introduction

COVID-19 is a Data Science issue" (Callaghan, 2020) the comprehensive article gives various ideas and inspiration to think about the data and how it can be effectively used in current pandemic situation. Quarantine is nothing but the separation and restriction of movement or activities of persons who are not ill but who are believed to have been visible to infection, for the purpose of avoiding transmission of diseases. People are usually quarantined in their homes, but they may also be quarantined in community-based accommodations. Considering the increasing volume of number of patients and limited community-based facilities most of the people are being asked to quarantine in their homes.

According to Wollersheim, 2020 during the COVID-19 crisis the field of Data Science is in epicenter. Almost whole public is interested, watching and looking forward for the statistical analysis and epidemiology graphs and sharing the same in social media on a large scale. The probability from Data Science is very high. Data Science is a developing field consists of number of appropriate and useful tools, functions and techniques.

Singh *et.al.* 2018, suggested the cluster containment strategy for Zika virus outbreak was found effective in Rajasthan, India. It explained how surveillance strategies used to control the disease from spreading beyond containment zones of 3 km radius. The article gives importance on creating containments to prevent the explosion of disease, however it does not explain about how to make these zones quickly and accurately. This paper (Maier & Brockmann, 2020) explains about the effective containment to control specifically COVID-19 cases in China. The model which they explained in their paper captures both quarantine of symptomatic infected individuals and other population isolation practices. The emphasis of the research is on infection process and general effects as well as significance of the containment. Their research work implies and supports the need to define the containment zones accurately.

Manimannan G. *et.al* (2021), predicts and classifies the data of COVID-19 based on four machine learning algorithm with four major parameters namely confirmed cases, recoveries, deaths and active cases. The secondary sources of database were collected from Ministry of Health and Family Welfare Department (MHFWD), from Indian State and Union Territories up to March, 2021. Based on these background, the database classified and predicted various machine learning Algorithm, like SVM, kNN, Random Forest and Logistic Regression. Initially, k-means clustering analysis is used to perform and identified five meaningful clusters and is labeled as Very Low, Low, Moderate, High and Very High of four major parameters based on their average values. In addition five clusters are cross validated using four machine algorithm and affected states are visualized in the table with help of prediction and probabilities. The different machine learning models cross validation and classification accuracy are 88%, 97%, 91% and 91%. The Classification of States and Union Territories were named as Very Low Affected (VLA), Low Affected (LA), Moderately Affected (MA), Highly Affected (HA) and Very Highly Affected (VHA) States and Union Territories of India by COVID-19 cases. Maharashtra is correctly classified as Very High Affected States, Delhi, Uttar Pradesh and West Bengal falls in Moderately Affected States, Assam, Bihar, Chattisgarh, Haryana, Gujarat, Madhya Pradesh, Odisha, Punjab, Rajasthan and Telangana falls in Low Affected States and Tamilnadu, Kerala Andhra Pradesh and Karnataka forms a group of highly affected States. Remaining States and Union Territories falls in Very Low affected by Covid-19 Cases.

COVID-19 Cases and Methods

Data science generally has a five-stage lifecycle that consists of Data science generally has a five-stage lifecycle that consists of:

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

Data Collection

The secondary sources of database were collected from Kaggle.com website, from April 2020 to April 2021. The original data consists of Total Cases, Positive and Negative cases of Indian states and union territories. Total Cases, Positive and Negative cases are the three parameters used in this research paper.

General Algorithm of Clustering Methods

The clustering techniques proceeded by either a series of successive mergers or a series of successive divisions. The following steps describe the hierarchical clustering algorithm for grouping N objects of parameters (R A. Johnson and D.W. Wichern, 2009).

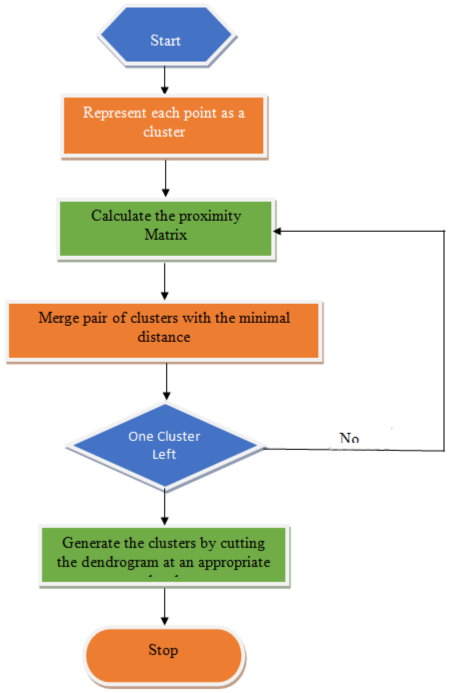
Step 1: Starts with N groups, each containing a single entity and $N * N$ symmetric of distances and is denoted by $D = \{d_{ik}\}$

Step 2: To identify the distance matrix for the nearest pair of groups, the distance between most similar groups X and Y be d_{xy}.

Step 3: Merge groups X and Y, Label the newly formed group (XY). Revise the entire database in the distance matrix by (a) deleting rows and columns of the corresponding cluster X and Y and (b) adding rows and columns giving the distances between the cluster (XY) and the remaining clusters.

Step 4: Repeat steps 2 and 3 a total of $N - 1$. (Figure 1.)

Figure 1. Flow Chart of Agglomerative Approach


Start
Represent each point as a cluster
Calculate the proximity Matrix
Merge pair of clusters with the minimal distance
One Cluster Left
No
Generate the clusters by cutting the dendrogram at an appropriate
Stop

Single Linkage Method

Step 1: The input to a single linkage method algorithm can be distance or similarities between pairs of items.

Step 2: Clusters are formed from individual entities by merging nearest neighbours, where the term nearest neighbour connects the smallest distance or largest distance similarity.

Step 3: Initially, to find smallest distance in $D = \{d_{ik}\}$ and merge the corresponding objects, X and Y, to get the cluster (XY). For step 3 of the general algorithm of hierarchical clustering method, the distances between XY and any other cluster W are computed by

$$d_{(XY)W}=\min\{d_{XW}, d_{YW}\}$$

where d_{XW} and d_{YW} are the distance between the nearest neighbours of clusters X and W and clusters Y and W respectively.

Step 4: The results of single linkage clustering can be graphically displayed in the form of a dendrogram.

Complete Linkage Method

Step 1: Complete linkage clustering algorithm similar to single linkage clustering's with one important exception.

Step 2: Every stage, the distance between clusters is determined by similarities between two elements one from each cluster that are most distant.

Step 3: Complete linkage ensures that all items in a cluster are within some maximum distance of each other.

Step 3: The general algorithm hierarchical clustering again starts by finding the maximum entry in $D = \{d_{ik}\}$ and merging the corresponding objects, such as X and Y, to get the cluster (XY).

Ste 4: For step 3 of the general algorithm, the distances between XY and any other cluster W are computed by

$$d_{(XY)W} = \max\{d_{XW}, d_{YW}\}$$

Where d_{XW} and d_{YW} are the distance between the nearest neighbours of clusters X and W and clusters Y and W respectively.

Step 4: The results of complete linkage clustering can be graphically displayed in the forms of a dendrogram.

Average Linkage Method

Step 1: Average linkage method treats the distance between two clusters as the average distance between all pairs of parameters where one number of a pair belongs to each other.

Step 2: Repeat, the input to average linkage algorithm may be distances or similarities, and the model can be used to group parameters or variables.

Step 3: Step 3 of the above general algorithm in the distance between (XY) and any other cluster W are computed by

$$. \, d_{(xy)w} = \frac{\sum_i \sum_k d_{ik}}{N_{xy} N_w}$$

Where d_{ik} is the distance i in the cluster between (XY) and the object k in the cluster W and N_{xy} and N_w are the number of items in clusters (XY) and W respectively.

Wards Method

Ward's hierarchical clustering algorithms based on minimizing the loss of information from joining two groups.

Step 1: This method is usually implemented with loss of information taken to be an increase in an error sum of squares criterion. ESS, first for a given cluster k, let ESS, be the sum of squared deviations of every item in cluster from cluster mean (centroid).

Step 2: If there are currently k clusters, define ESS as the sum of the $\mathrm{ESS_k}$ or $ESS = ESS_1 + ESS_2 + ESS_3 + \cdots + ESS_k$.

Step 3: At each step in analysis, the union of every possible pair of clusters is considered, and the two clusters whose combination results in the smallest increase in ESS are joined.

Step 4: Initially, each cluster consists of a single item, and, if there are N items $ESS_k = 0, k = 1,2, \dots, N$, so $ESS = 0$.

Step 5: At the other extreme, when all clusters are combined in a single group of N terms, the value of ESS is given by

$$ESS = \sum_{j=1}^{N} (x_j - \bar{x})'(x_j - \bar{x})$$

Where, x_j is the multivariate measurement associated with the jth item and $\bar{x}$ is mean of all the items. The results of Ward's method can be displayed as a dendrogram (Ward. Jr, 1963).

Euclidean Distance Similarity

Most machine learning algorithms including Hierarchical Cluster use this distance metric to measure the similarity between observations.

Here's the formula for Euclidean Distance:

$$d = ((p_1 - q_1)^2 + ((p_2 - q_2)^{1/2}$$

We use this formula when we are dealing with 2 dimensions. We can generalize this for an n-dimensional space as:

$$D_e = \left(\sum_{i=1}^{n} (p_i - q_i)^2\right)^2$$

Where,

- $n =$ number of dimensions
- $p_i, q_i =$ data points

The above formula contains both procedures and functions to calculate similarity between sets of data. The function is best used when calculating the similarity between small numbers of sets. The procedures parallelize the computation and are therefore more appropriate for computing similarities on bigger datasets (Ian H. Witten *et. al. (2016).*

COVID-19 Model Developing Algorithm

Step 1: Covid-19 database of Indian States and Union Territories is given as input data matrix to file widget.

Step 2: To view the data matrix from data table, widget to be connected in file widget.

Step 3: The Cosine distance matrix is calculated by distance widget, after that it is connected to distance matrix widget, distance map widget and Hierarchical clustering widget.

Step 4: The distance map, distance matrix widgets are calculated using the distance matrix and display the distance map of Indian States and Union Territories database.

Step 5: The hierarchical clustering widget is used to calculate various methods of clustering and visualize the results in dendrogram.(Figure 2) and the formation of clusters can be viewed from data table (1) widget.

Step 6: Each method of clustering results are interpreted in result and discussion section.

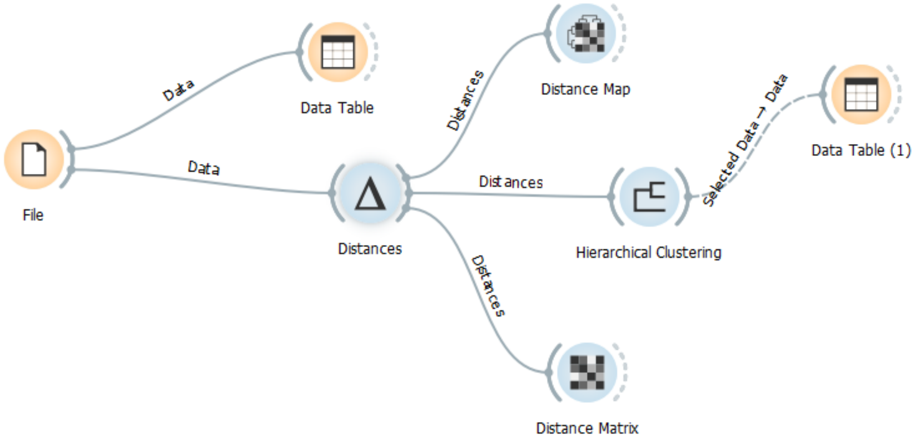
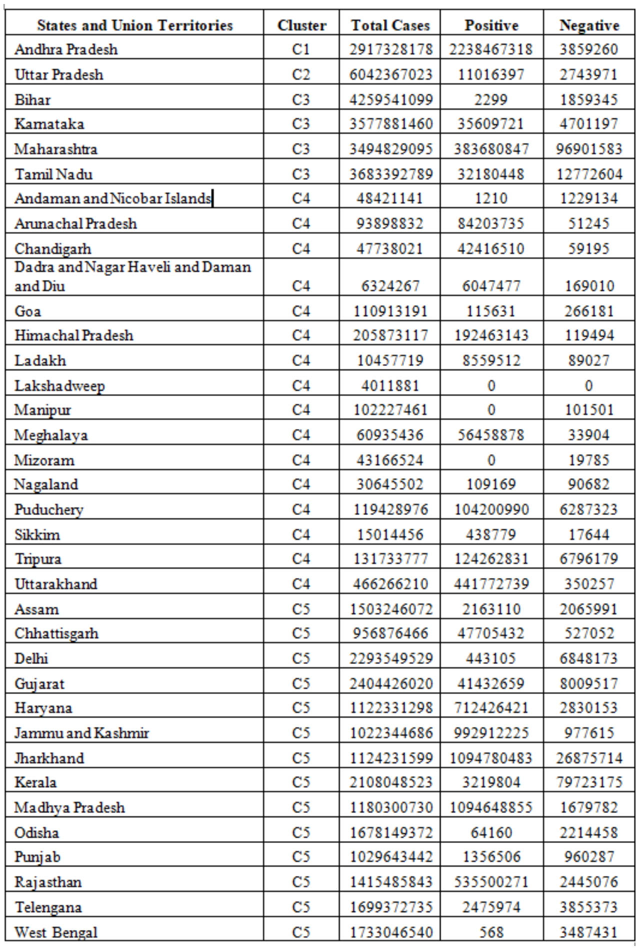


Figure 2. Workflow and Widget of Hierarchical Clustering Technique

Table 1: Formation of Clustering Methods and their Rsults

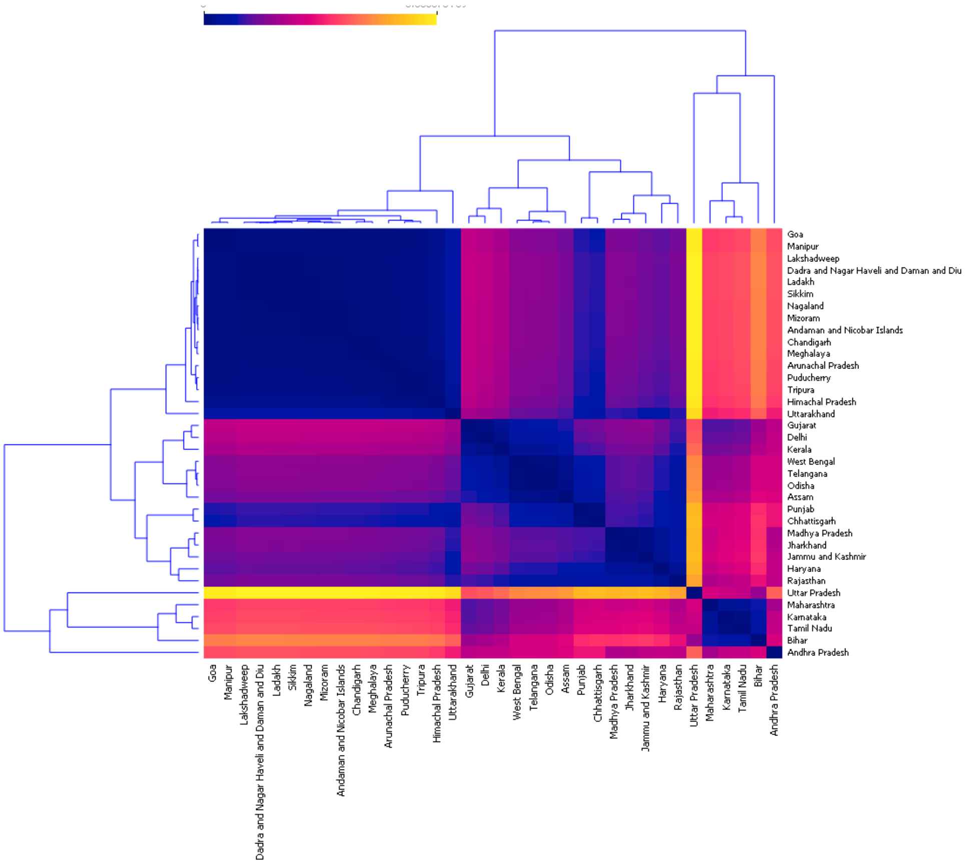


States and Union Territories	Cluster	Total Cases	Positive	Negative
Andhra Pradesh	C1	2917328178	2238467318	3859260
Uttar Pradesh	C2	6042367023	11016397	2743971
Bihar	C3	4259541099	2299	1859345
Kamataka	C3	3577881460	35609721	4701197
Maharashtra	C3	3494829095	383680847	96901583
Tamil Nadu	C3	3683392789	32180448	12772604
Andaman and Nicobar Islands	C4	48421141	1210	1229134
Arunachal Pradesh	C4	93898832	84203735	51245
Chandigarh	C4	47738021	42416510	59195
Dadra and Nagar Haveli and Daman and Diu	C4	6324267	6047477	169010
Goa	C4	110913191	115631	266181
Himachal Pradesh	C4	205873117	192463143	119494
Ladakh	C4	10457719	8559512	89027
Lakshadweep	C4	4011881	0	0
Manipur	C4	102227461	0	101501
Meghalaya	C4	60935436	56458878	33904
Mizoram	C4	43166524	0	19785
Nagaland	C4	30645502	109169	90682
Puduchery	C4	119428976	104200990	6287323
Sikkim	C4	15014456	438779	17644
Tripura	C4	131733777	124262831	6796179
Uttarakhand	C4	466266210	441772739	350257
Assam	C5	1503246072	2163110	2065991
Chhattisgarh	C5	956876466	47705432	527052
Delhi	C5	2293549529	443105	6848173
Gujarat	C5	2404426020	41432659	8009517
Haryana	C5	1122331298	712426421	2830153
Jammu and Kashmir	C5	1022344686	992912225	977615
Jharkhand	C5	1124231599	1094780483	26875714
Kerala	C5	2108048523	3219804	79723175
Madhya Pradesh	C5	1180300730	1094648855	1679782
Odisha	C5	1678149372	64160	2214458
Punjab	C5	1029643442	1356506	960287
Rajasthan	C5	1415485843	535500271	2445076
Telengana	C5	1699372735	2475974	3855373
West Bengal	C5	1733046540	568	3487431

Results and Discussion

From the above workflow, visualization map and distance matrix has been computed. The maximum distance matrix similarity value is represented in dark colors and minimum distance matrix similarity value is highlighted in light shade and a part of distance matrix in Distance Map in Figure 3.

Figure 3 Distance Map



From the following figure (Figure 4 to 7), it is very clear that 5 zones of C1, C2, C3, C4 and C5 have formed. The C1 zone indicates Uttar Pradesh state has Very High (VH) total, positive and negative cases of cluster, C2 zone indicates Bihar, Karnataka, Maharashtra and Tamilnadu states have High (H) total, positive and negative cases. C3 zone indicates state Andhra Pradesh has Low (L) total, positive and negative cases.

C4 zone indicates that the states and union territories of Andaman and Nicobar Islands, Arunachal Pradesh, Chandigarh, Dadra and Nagar Haveli and Daman and Diu, Goa, Himachal Pradesh, Ladakh, Lakshadweep, Manipur, Meghalaya, Mizoram, Nagaland, Puduchery, Sikkim, Tripura and Uttarakhand have recorded Very Low (VL) Total, Positive and Negative cases.

The final Cluster C5 zone indicates that the states and union territories of Assam, Chhattisgarh, Delhi, Gujarat, Haryana, Jammu and Kashmir, Jharkhand, Kerala, Madhya Pradesh, Odisha, Punjab, Rajasthan, Telengana and West Bengal has reported a Moderate number of Total, Positive and Negative cases. In all the clustering methods five clusters have formed with same States and Union Territories. This result indicates all machine learning cluttering methods formed as natural clusters using Euclidean Distance with three parameters on Indian States and union Territories.

Figure 4. Single Linkage

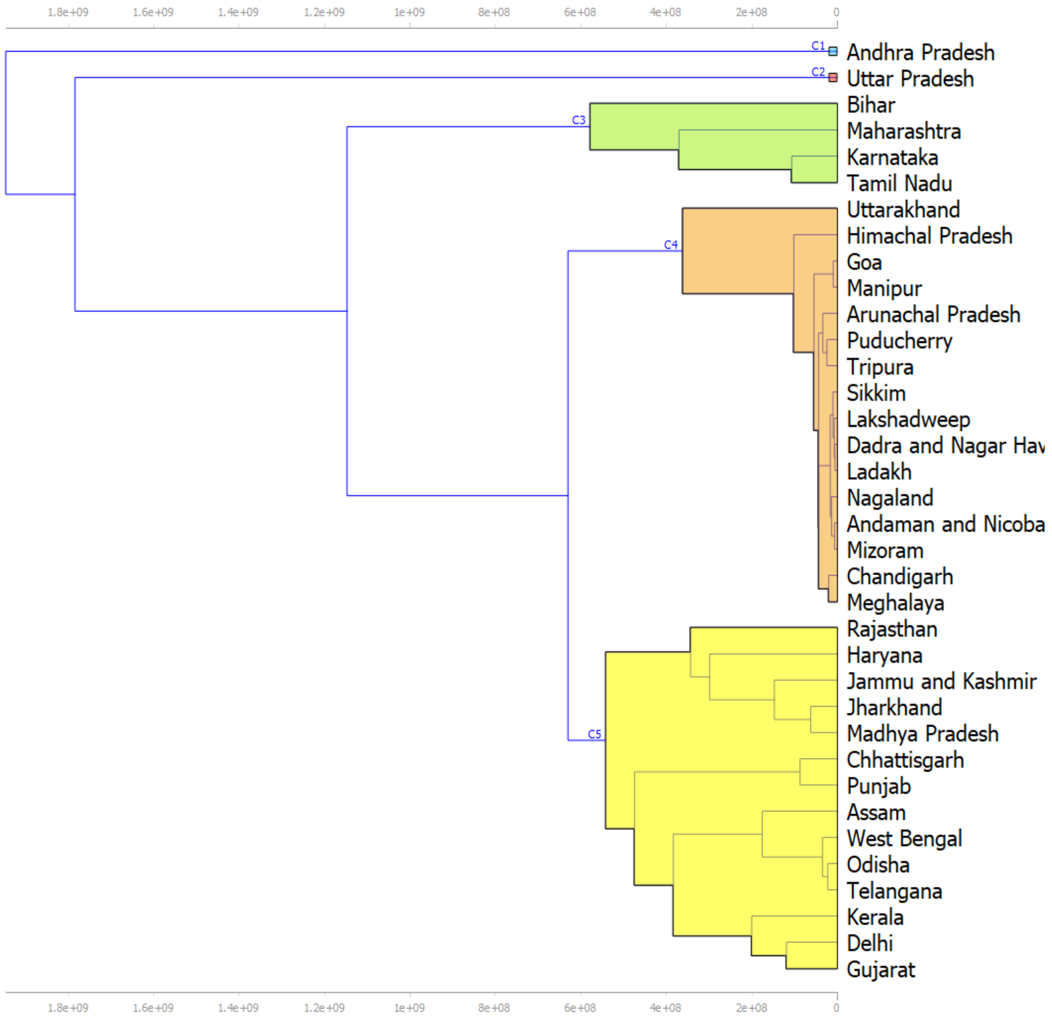


Figure 5. Average Linkage

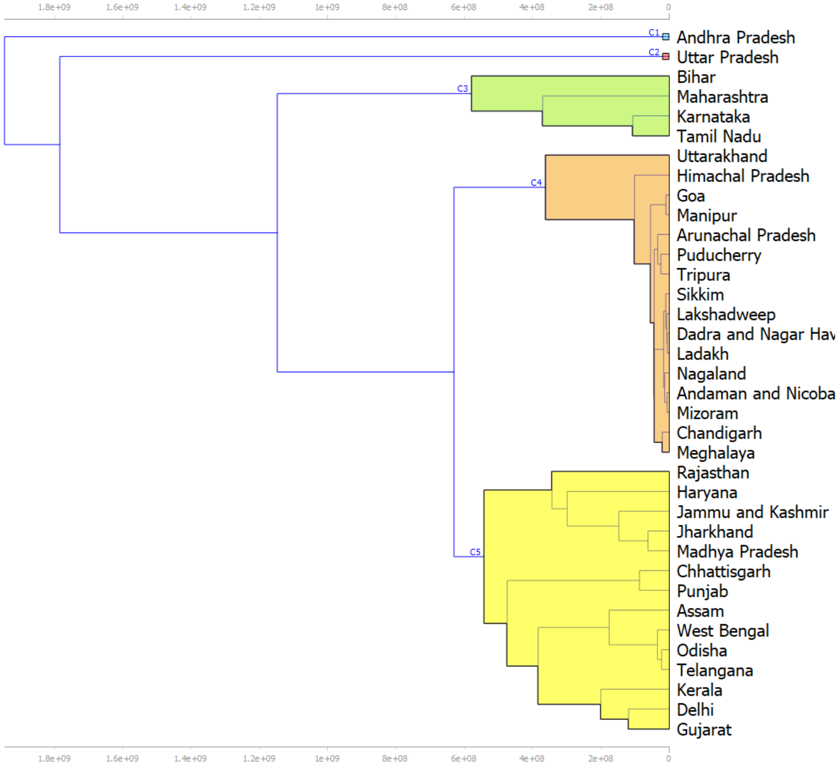


Figure 6. Complete Linkage

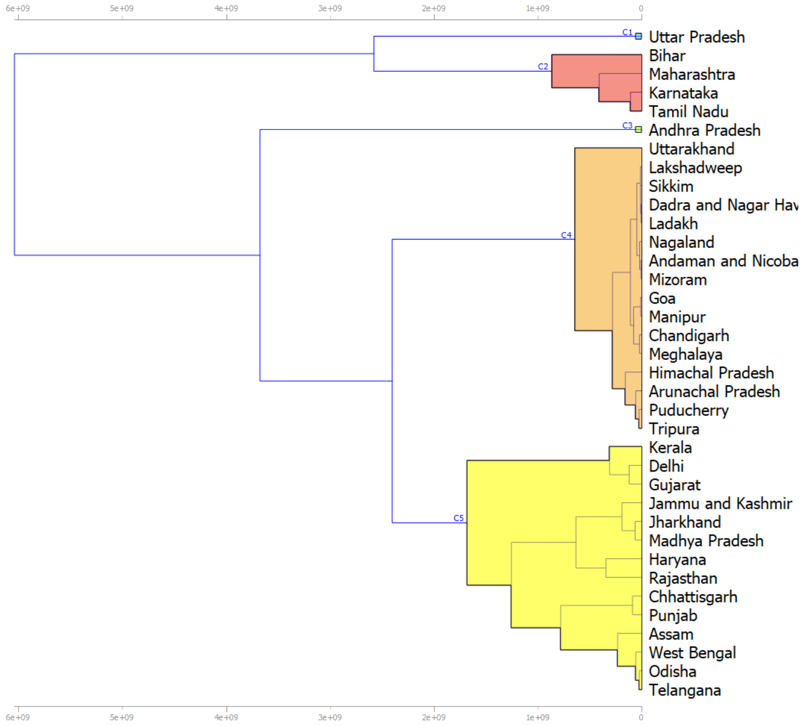
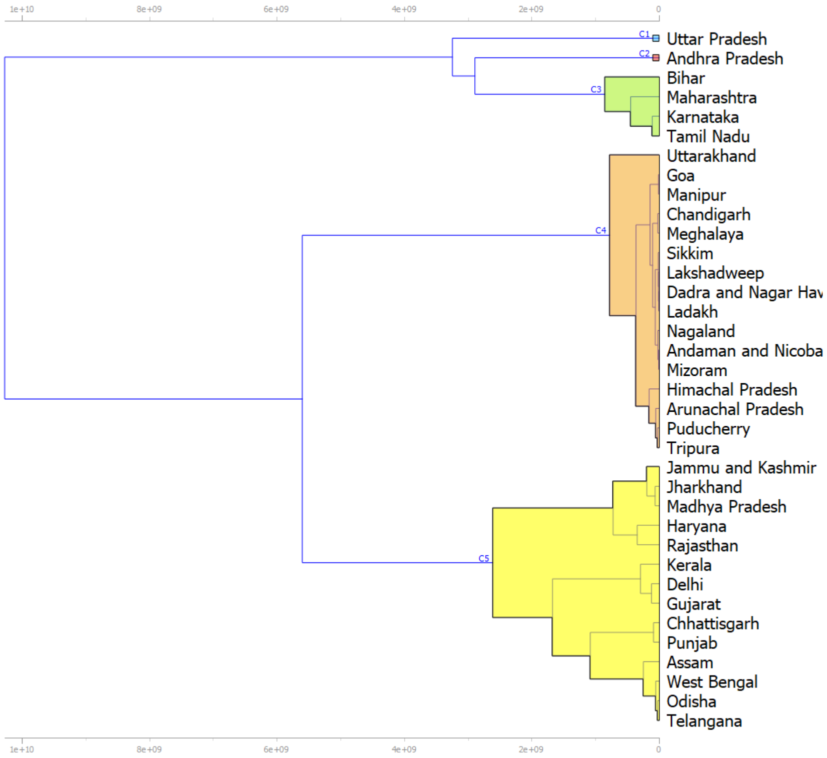


Figure 7. Ward's Linkage Method



All clustering methods are visualized in the dendrogram (Figure 4 to 7) and separated by five different clusters of Indian states and union Territories. C1 Very highly affected state is Uttar Pradesh and C4 zone of States and union Territories are very Low affected of total, positive and negative cases.

Conclusions

Application of Orange Data mining software determines the various hierarchical clustering methods and visualized the results through dendrogram using the parameters Total, Positive and Negative cases of sample data from various states and union territories. The Secondary sources of data were collected from April 2020 to April 2021 with the help of three main parameters: Total Cases, Negative Cases and Positive Cases. Subsequently the python based Orange data mining workflow executed the hierarchical clustering methods of Single Linkage, Complete Linkage, Weighted Linkage, Average Linkage and Wards method. The file widget open new COVID-19 database and perform hierarchical cluster with Euclidean distance measure. The Euclidean distance measure achieved five natural clusters.

The five clusters are visualized in the form of dendrogram and showed that results of Indian states and union territories. They are labeled as five different clusters and are labeled as C1, C2, C3, C4 and C5. The C1 zone indicates that Uttar Pradesh has recorded Very High (VH) total, positive and negative cases of cluster; C2 zone indicates that Bihar, Karnataka, Maharashtra and Tamilnadu states has recorded High (H) total, positive and negative cases.

C3 zone indicates Andhra Pradesh has recorded Low (L) total, positive and negative cases. C4 zone indicates that the states and union territories of Andaman and Nicobar Islands, Arunachal Pradesh, Chandigarh, Dadra and Nagar Haveli and Daman and Diu, Goa, Himachal Pradesh, Ladakh, Lakshadweep, Manipur, Meghalaya, Mizoram, Nagaland, Puduchery, Sikkim, Tripura and Uttarakhand has reported Very Low (VL) Total, Positive and Negative cases.

The final Cluster C5 zone indicates that the states and union territories of Assam, Chhattisgarh, Delhi, Gujarat, Haryana, Jammu and Kashmir, Jharkhand, Kerala, Madhya Pradesh, Odisha, Punjab, Rajasthan, Telengana and West Bengal has reported a Moderate number of Total, Positive and Negative cases.

References

1. Callaghan, S. (2020). COVID-19 Is a Data Science Issue. In *Patterns* (Vol. 1, Issue 2, p. 100022). https://doi.org/10.1016/j.patter.2020.100022.
2. Graeme Wetherill and Paul W. Burton, (2008), Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region, Geophysical Journal International 176(2):565 - 588
3. Maier, B. F., & Brockmann, D. (2020). Effective containment explains sub exponential growth in recent confirmed COVID-19 cases in China. *Science*, *368*(6492), 742–746. https://doi.org/10.1126/science.abb4557
4. Manimannan G. et.al (2021), Prediction, Cross Validation and Classification in the Presence COVID-19 of Indian States and Union Territories using Machine Learning Algorithms, International Journal of Recent Technology and Engineering (IJRTE Volume 10 issue 1, pp. 16-20
5. R A. Johnson and D.W. Wichern (2009), Applied Multivariate Statistical Analysis, PHI Learning Private Limited, India.
6. Ward. Jr. T. H, (1963), Hierarchical Grouping to optimize as Objective Function. Journal of The American Statistical Association, 58, pp.236-244.
7. Ian H. Witten et. al. (2016). Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann; 4th edition