Social Network Analysis and Mining (2021) 11:28
https://doi.org/10.1007/s13278-021-00734-2

REVIEW PAPER

Community detection using unsupervised machine learning techniques on COVID-19 dataset

Laxmi Chaudhary[1] · **Buddha Singh**[1]

Received: 31 August 2020 / Revised: 30 December 2020 / Accepted: 12 February 2021 / Published online: 10 March 2021

Abstract
COVID-19 has been considered to be the most destructive pandemic ever happened in the history of mankind. The worldwide research community has put a tenacious effort to carry out research on the COVID-19 to analyse its impact on economic, medical and sociolgoical fields. They are trying to solve many crucial issues related to this disease and derive strategies to deal with this global pandemic. In this paper, we have analysed the trend, countries affected regionally and the variation of cases at the country level on COVID-19 dataset. We have used the Principal component analysis on the COVID-19 dataset variables to reduce the dimensionality and find the most significant variables. Further, we have unveiled the hidden community structure of countries by applying the unsupervised clustering approach, K-means. We have compared the results with the K-means method. The communities achieved after applying the PCA are more precise. The resulted communities can be beneficial to researchers, scientists, sociologists, different policy makers and managers of health sector.

Keywords COVID-19 · Coronavirus · K-means · PCA · Communities · Machine learning

1 Introduction

The coronavirus COVID-19 is rapidly spreading across the world since the beginning of 2020. WHO (World Health Organization) has categorised it as a global pandemic (WHO. Briefing by WHO Director-General Tedros Adhanom Ghebreyesus.March 11, 2020 (Accessed at:https www.pscp.tv,1djxXQkqApVKZ). 2020), due to its highly contagious nature. In the current global pandemic situation, all the countries are struggling with COVID-19 and still looking for a cost-effective and practical solution to encounter the challenges arising in various ways. Researchers from different fields such as engineering, physical and medical sciences are attempting to take such challenges, to develop new theories, and to generate user-centered solutions (Singh et al. 2020).

Recent studies on COVID-19 have mainly focused on the analysis at individual level, that is based on its attributes and symptoms of this disease. The studies on the various geographical areas and huge populations have been inconsequential till now. Hence, there is a significant scope for research in this area other than the research done on patients information (Carrillo-Larco and Castillo-Cara 2020).

In this paper, the analysis at the country-level COVID-19 dataset could provide potentially modifiable related factors that indiviual level dataset is not able to uncover due to the limited variables. Furthermore, the analytical approach such as unsupervised machine learning for community detection is used to analyse the behavior of countries during the COVID-19 global pandemic. Hence, the community detection method helps in unveiling the patterns of countries and regions where the COVID-19 has impacted in a similar pattern. Regions and countries could use this pattern and information to prohibit worst situation.WHO and other global organization could use this information to give countries a similar aid. Therefore, we have developed an unsupervised machine learrning method, PCA (Principal component analysis) (Shlens 1404) and K-means clustering (Figueiredo and Jain 2002) on the country level COVID-19 pandemic dataset, that can detect communties of countries regarding the country level variables. It means, we aimed to find out: the significant variables of the COVID-19 pandemic by applying the PCA. The PCA helps in reducing the dimensions of the COVID-19 data without loosing any information and

✉ Laxmi Chaudhary
laxmichaudari.iet@gmail.com

Buddha Singh
b.singh.jnu@gmail.com

[1] Jawaharlal Nehru University, New Delhi 110067, India

provides the most significant characteristics. Furthermore, we have detected the communities of countries using the reduced dataset efficiently.

In this research paper, our contribution are as follows:

-The analysis of the country level COVID-19 dataset, which helps in understanding the count of the countries and affected in various WHO regions.

-The analysis of the top four most affected countries gives better idea how the various cases are varying at country-level.

-The PCA helps in finding the most significant patterns in the data that helps in reducing the number of dimensions without losing much of the information.

-We have compared the results of K-means after applying PCA with the results of the K-means method.

-The community detection approach grouped the countries, that helps in objectively distinguish countries and regions with respect to the COVID-19 dataset spread and results.

-Different policy makers, for instance managers, physicians from the health sector, researchers and scientists to make use of these analyses.

2 COVID-19 dataset description and analysis

The COVID-19 dataset is collected from the official website of Johns Hopkins University (Marutho et al. 2018). It consists the number of cases from January 22nd, 2020 to August 15th, 2020. The Excel 2019 has been used to collect and integrate the dataset. The final dataset can be retrieved from [7]. The retrieved country level data is recorded into excel and further analysed. The country level dataset has the data of 187 countries and the number of variables is 15.

Figure 1. shows the number of countries impacted due to COVID-19 with respect to the WHO regions. We can see that a total of 56 countries in European and 48 countries in African region are highly affected due to coronavirus. Figure 2. shows the percentage of total deaths due to COVID-19 corresponding to the WHO regions. It indicates that the American regions have been most affected with 54.16% deaths followed by European regions with 28.45%. The total death percentage is less in African regions inspite of having second most affected region with 48 countries. Moreover, the South-east asia region has ten countries impacted due to COVID-19 which is the lowest as compared to other regions, however, it has higher death percentage than the Western Pacific region which has 16 countries affected.

Figure 3. shows the percentage of cases in the Top 4 most affected countries. It provides the percentage of confirmed, deaths, recovered and active cases from Brazil, US, India and Russia. The US is the hardest hit country due to the COVID-19 pandemic. The total death percentage is least in Russia as compared to other countries.

3 Unsupervised machine learning for community detection on COVID-19

In this section, we have described the Principal components method and the K-means method to identify the communities of countries affected from COVID-19 dataset based on the cases. For implementation and visualization of the dataset, we have used Python 3.7 and Tableau 2019.

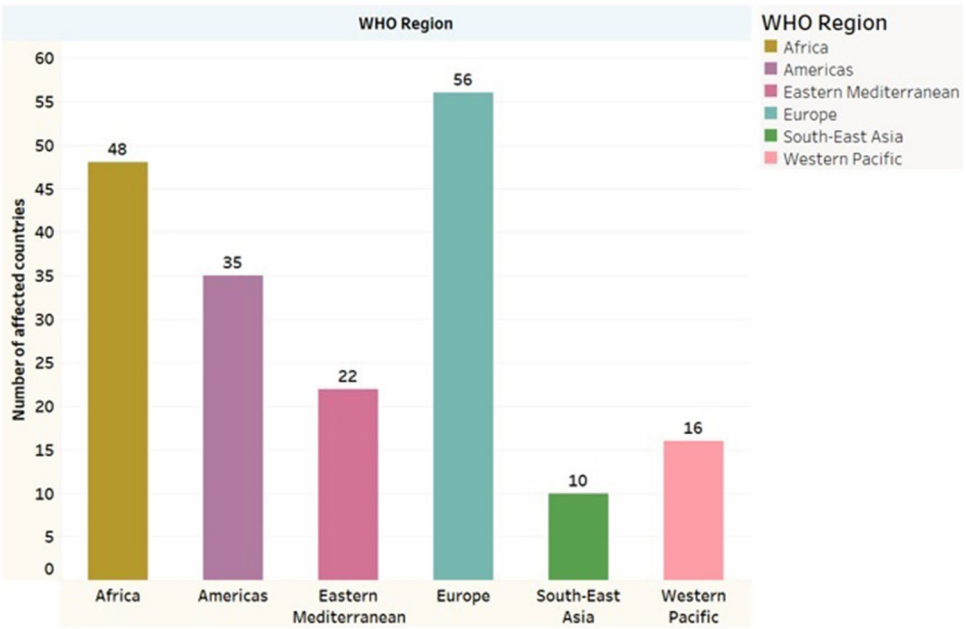

Fig. 1 Number of countries impacted due to COVID-19 w.r.t the WHO regions

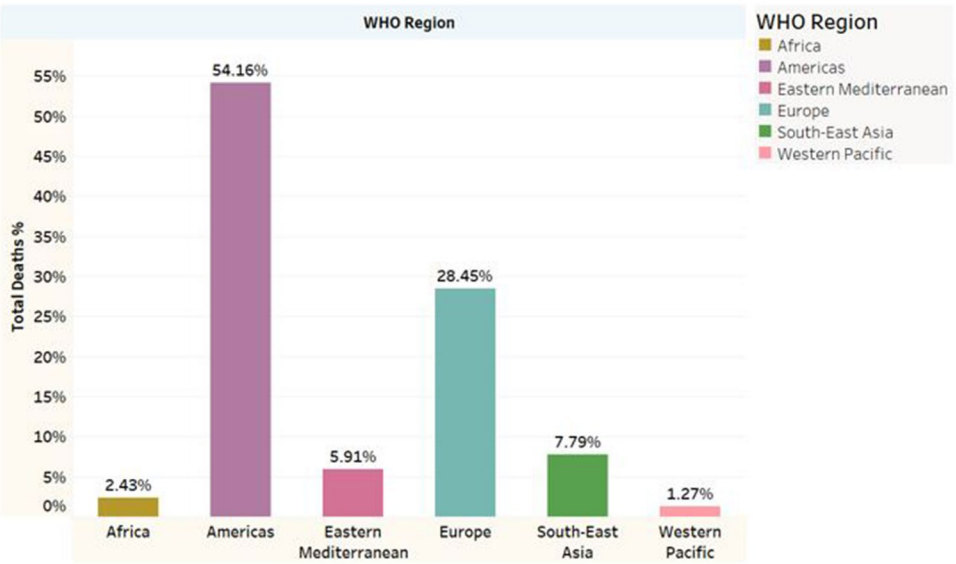

Fig. 2 Percentage of total deaths due to COVID-19 w.r.t the WHO regions

Fig. 3 Percentage of cases in the Top 4 most affected countries

4 Dimensionality Reduction using PCA

The Principal component analysis (PCA) is an essential approach for the pattern analysis of the data. After finding the patterns, it reduces the dimensions of the data without loosing much of the information. PCA maximizes the variance of the projected data and minimizes the squared error between datapoints and their projections.The COVID-19 data consists of 13 continuous variables and some of these variables are highly correlated. Due to the orthogonal transformation, the PCA turns strongly related variables into uncorrelated variables. The PCA has been helpful in creating characteristics set, that illustrates the related information from the COVID-19 data of 13 variables. This is a representation of the number of variable reductions and maximization of the variance.

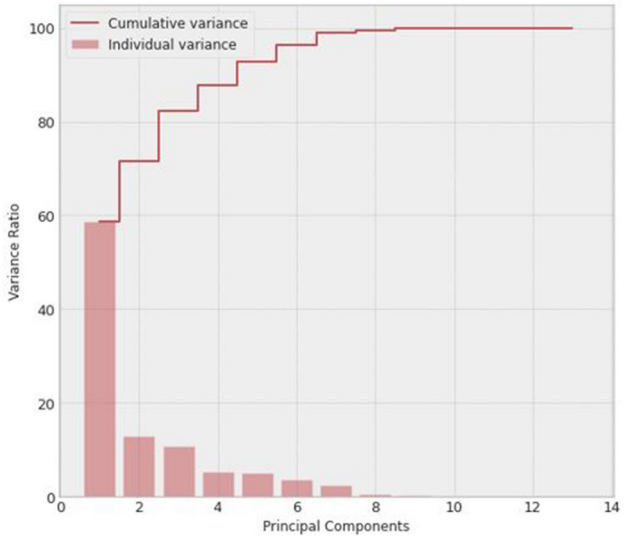

Fig. 4 Cumulative variance of the principal components

Given the COVID-19 dataset, we start by standardizing the data to create the clustering model and the variables need to have other values between them, hence each one of them have a different variance. Due to this fact, it is significant to normalize these variables to find reliable communities with retaining relevant information. Then, we have to generate the covariance matrix for all the variables. Covariance can be computed using formula,

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{(n-1)} \quad (1)$$

where X_i is the data points of the variable, n is the number of datapoints in the dataset and $\overline{X}$ is the mean and is given by $\overline{X} = \sum_{i=1}^{n} X_i/n$. Similarly, Y_i is defined. Using the covariance values, we have constructed the covariance matrix, From the covariance matrix, we have calculated the eigenvalues and eigenvectors. Thus, according to this process of considering the covariance matrix' eigenvectors, we have been able to get the lines that characterizes the data. The eigenvector with the largest eigenvalues is the principal components of the dataset. So, we order the eigenvalues in descending order. It gives siginificance order of principal components. Figure 4. shows the cumulative variance of the principal components. The eight principal components have contributed the 99.9 cumulative variance. The eight PCA components preserved a variance of almost 1. Hence, we have taken eight principal components and ignored the other components of lesser significance. This approach of acquiring 100% as an explained variance signifies that retaining 100% of the information explained by the original 13 variables. It clearly indicates that 8 variables are sufficient to detect the communities instead of 13 variables. Additionally, these eight components provide the most accurate communities as explained in the next section. The steps of the PCA Algorithm are described in Algorithm1.

Algorithm1: PCA Algorithm		
Input	:	COVID-19 dataset, K (number of clusters)
1	:	Standardize the dataset.
2	:	Generate the covariance matrix using the covariance values using (1).
3	:	Compute eigenvalues (which are the magnitudes of variance captured) and eigenvectors (which are the principal components).
4	:	Sort the eigen pairs in descending order of eigenvalues and select the largest eigenvalues components which captures the variance of 1.
Output	:	Reduced new dataset.

4.1 Community detection using centroid based K-means approach

Now after obtaining the dataset, we have applied unsupervised clustering method, K-means on reduced dataset. K-means method uncovers the communities from the heterogeneous elements and clusters them into homogeneous groups. It groups the elements into clusters that were undefined at the beginning of the analysis. This methodology has been used earlier in various sectors such as clinical and public health research. Different methods of unsupervised clustering depend on the characteristics of the dataset. In this research work, we have taken centroid based K-means algorithm. It is suitable for communities that are in similar densities, similar size, and have a globular shape.

K-means method requires the information about K (number of clusters) as the input. Therefore, we have used the Elbow method [8] to identify the optimal number of clusters, K. The elbow method measures the homogeneity or heterogeneity within communities as the number of clusters

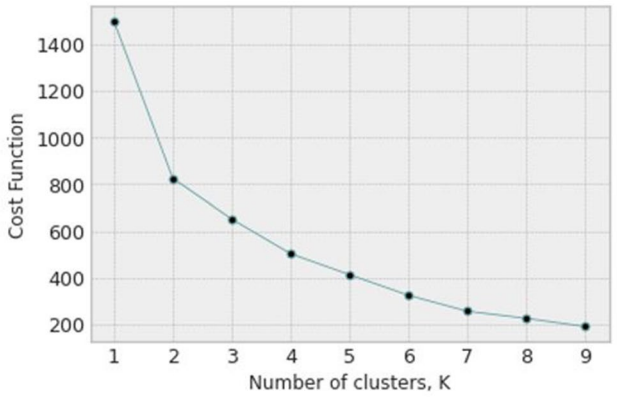

Fig. 5 Elbow function for K-means method

changes. Figure 5 shows the elbow function plot for the communities which maximizes the function convergence to the centroids. The dataset from the countries have been paired to support the selection of six and seven clusters. The selection of six and seven cluster of the countries in the group is based on the one who shares common socio-demographic and epidemiological profiles.

The number of clusters around the elbow function delivers almost similar information due to the limited number of observations done in this analysis. Visual analysis of maps and plots has been utilized to make a crucial decision of the number of communities to obtain the best output by grouping countries in a stable cluster of same background.

The retrospect analysis recommends to select six and seven clusters for K means method. During the visual inspection of the maps, geographical, epidemiological, and geopolitical knowledge have been used as input variables. Community detection has been done using six and seven clusters and they have given productive results based on the prior knowledge. Overall, our decision to go with six and seven clusters has been a good decision. Hence, we have detected the communities using K-means method after finalizing the number of clusters. The steps of the K-means method are described in Algorithm1.

Algorithm2: K means Method		
Input	:	Normalized reduced new dataset, K (number of clusters)
1	:	Choose, K (achieved from Elbow method) as initial centroid.
2	:	Construct K communities by assigning all data points to the nearest centroid.
3	:	Update centroids by calculating central data points of clusters.
4	:	Iterate above two steps until no datapoint is reassigned to another community.
Output	:	Communities of countries.

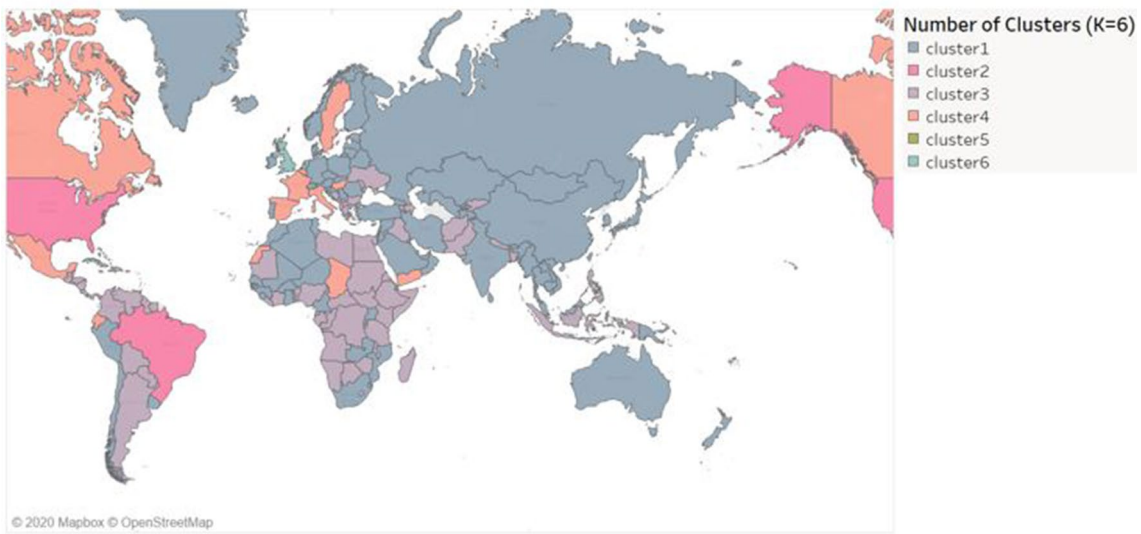

Fig. 6 Communities of countries using K-means ($K=6$) after applying PCA

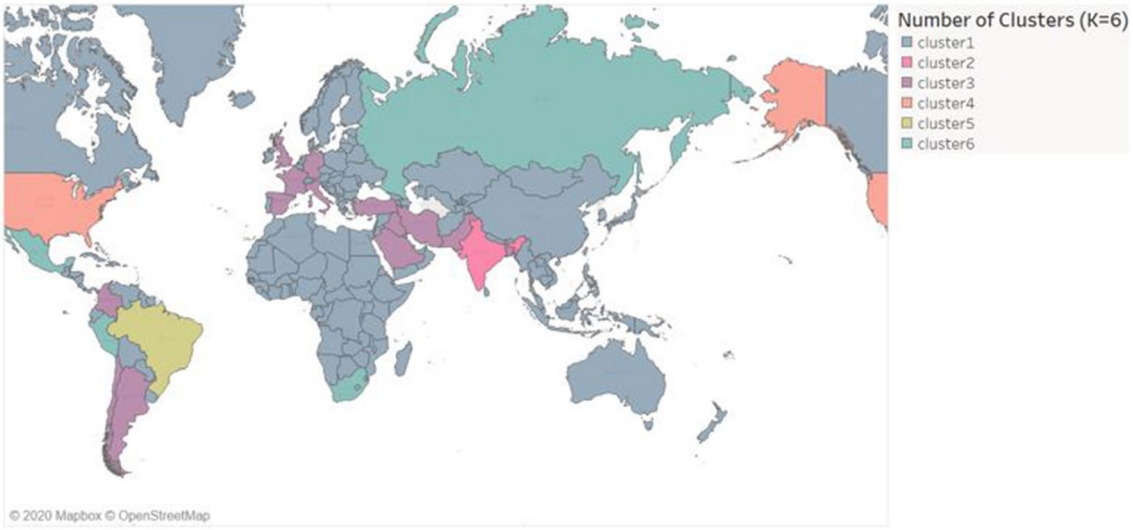

Fig. 7 Communities of countries using K-means ($K=6$) without applying PCA

5 Simulation results and comparison with the K-means method without PCA

In this section, we have depicted the simulation results of the method. We have analysed the results of K-means method after applying PCA and compared with the results of the K-means method. Figure 6. shows the communities obtained using the K-means ($K=6$) approach after applying PCA. It shows that US and Brazil in cluster2, Seychelles in cluster5, UK and Netherlands in cluster6 and rest of the countries in other clusters. Figure 7. shows the communities obtained using only the K-means ($K=6$) approach. It shows that India in cluster2, US in cluster4, Brazil in cluster 5, Russia, South Africa, Peru and Mexico in cluster6 and rest of the countries in other clusters.

We can clearly see that K-means after applying the PCA is giving better results as it has clustered US and Brazil in one cluster which is the most affected countries whereas K-means method has clustered both the countries separately.

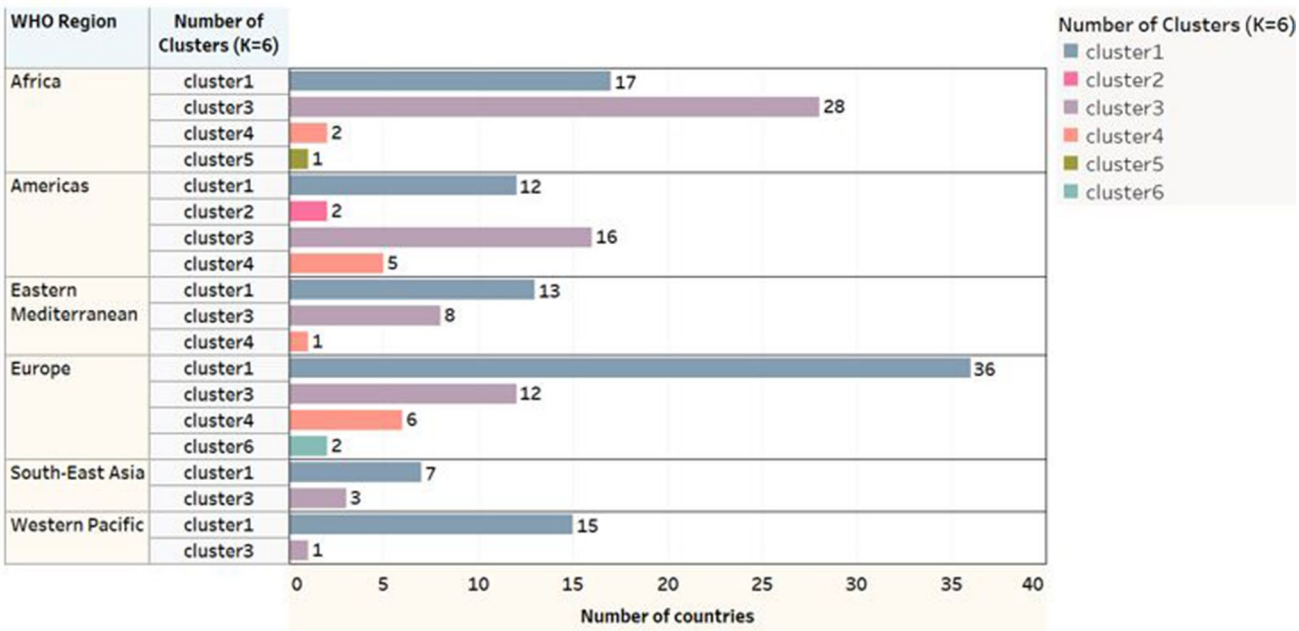

Fig. 8 Clusters of countries count region wise using K-means ($K=6$) after applying PCA

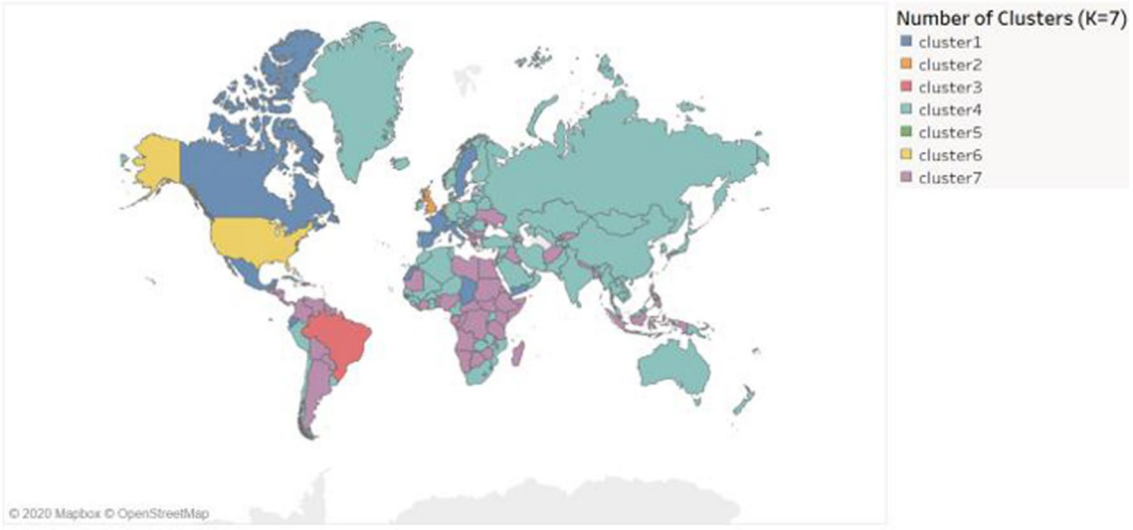

Fig. 9 Communities of countries using K means ($K=7$) after applying PCA

Fig. 10 Communities of countries using K-means ($K=7$) without applying PCA

Similar difference we can also observe for other clusters. Hence, the K-means method after applying the PCA is providing good communities based on the number of cases.

Figure 8. shows the count of the countries communities regionally using K-means ($K=6$) after applying PCA. It shows cluster1 and cluster3 countries is overlapping with all the regions due to COVID-19 cases. Cluster2 is belonging to Americans regions only (US and Brazil). Cluster4 is overlapping with four regions. Cluster6 is belonging to European region only.

Figure 9. shows the communities obtained using the K-means when $K=7$. It shows that US in cluster6, Seychelles in cluster5, Brazil in cluster3, UK and Netherlands cluster2 and rest of the countries in other clusters. Figure 10. shows the communities obtained using the K-means when $K=7$. It shows that India in cluster5, US in cluster3, Russia in cluster7, Brazil in cluster2, South Africa, Peru, Colombia, Mexico and Chile cluster6 and rest of the countries in other clusters.

We have observed that K-means after applying the PCA is providing better results as it has clustered India and Russia in one cluster which is less affected countries as compared to Brazil and US whereas K-means method has clustered both the countries separately. Similar trend we can also observe for other clusters. Therefore, the K-means method after

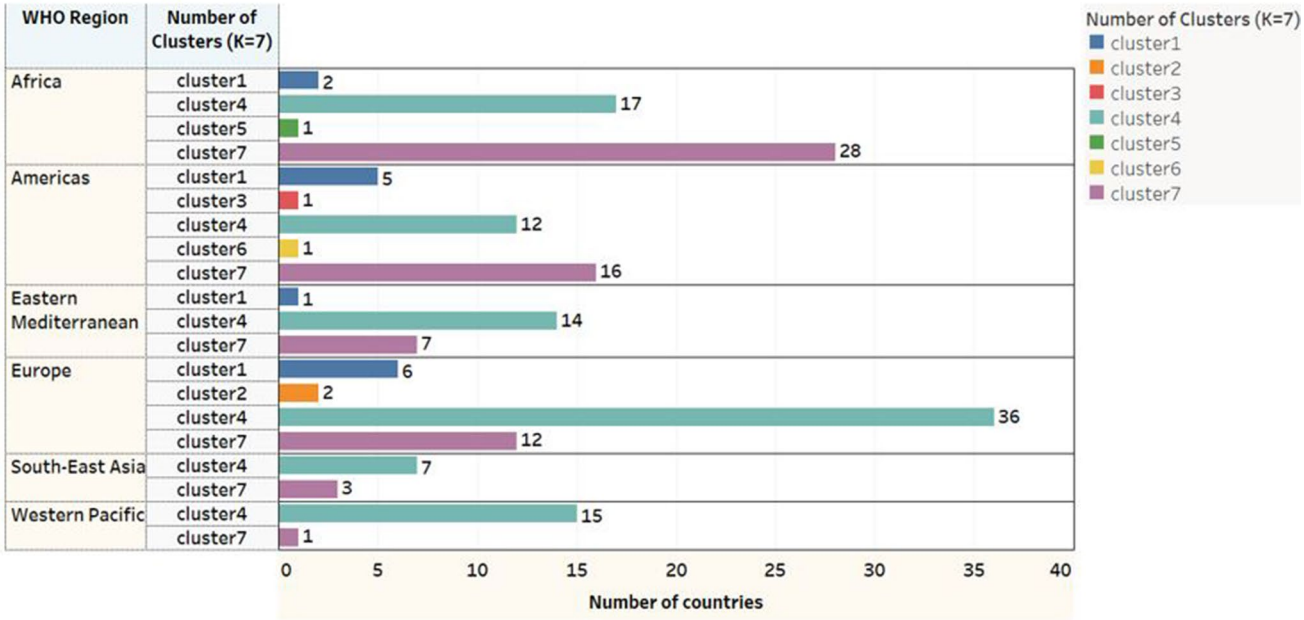

Fig. 11 Countries count region wise in communities using K-means ($K=7$) after applying PCA

applying the PCA is displaying good communities depending on the number of cases.

Figure 11. shows the count of the countries regionally in communities using K-means ($K=7$) after applying PCA. It shows cluster1 is overlapping with four regions. Cluster2 is belonging to European regions only (UK and Netherlands). Cluster4 and cluster7 countries is overlapping with all the regions. Cluster5 is belonging to African region. Cluster6 is belonging to American region.

6 Conclusion

In this research paper, we have analysed the trend of the countries affected regionally and also analysed the variation of cases at the country level on COVID-19 dataset. We have used the unsupervised machine learning approach, Principal component analysis on the COVID-19 dataset variables to reduce the dimensionality by covering the variance of 100% and find the most significant variables. Further, we have detected the hidden community structure of countries by applying the other unsupervised approach, K-means. We have compared the results of the K-means after applying PCA with the K-means method. The K-means after applying PCA is showing promising results. The communities of countries obtained can be beneficial in making various policies in health sector, also this information can help physicians and economy experts. It could also be helpful for countries and regions which belong to the same communities to provide similar aid also in taking preventive measures to avoid worst-case scenarios. In the future, this research work can be extended with more attributes to increase the number of variables, e.g., the overall population, climate, age distribution, etc. to see the variation in the cluster formation. Further, we will consider other dimensionality reduction methods like autoencoder, Forward/Backward feature construction, etc. to analyse the changes in the dimensions and its effect on the communities of countries.

Acknowledgement This research was funded by NFOBC fellowship of University Grants Commission under the Ministry of Human Resource Development (Government of India).

Reference

Carrillo-Larco RM, Castillo-Cara M (2020) Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. Wellcome Open Res 5(56):56

Figueiredo MAT, Jain AK (2002) Unsupervised learning of finite mixture models. IEEE Transac pattern analysis mach intell 24(3):381–396

Marutho, Dhendra, Sunarna Hendra Handaka, and Ekaprana Wijaya (2018) The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. 2018

International Seminar on Application for Technology of Information and Communication. IEEE
Shlens, Jonathon (2014). A tutorial on principal component analysis. arXiv preprint.
Singh, Ravi Pratap, et al. (2020) Internet of things (IoT) applications to fight against COVID-19 pandemic. Diabetes Metabolic Syndrome: Clinical Research Reviews.
WHO. Briefing by WHO Director-General Tedros Adhanom Ghebreyesus. March 11, 2020.(Accessed at: https://www.pscp.tv/w/1djxXQkqApVKZ).