Research Paper

Data Analysis of Some Neighbouring States In India: Covid-19 Pandemic

Dr. Vikas Kumar[1] and Dr. Ashish Arora[2]

[1]*Department of Mathematics and Statistics, Swami Vivekanand Subharti University*
[2]*SRS Institute of Education and Technology, Chaudhary Charan Singh University*

ABSTRACT
The development of contagious disease has a significant impact on world health, and a thorough assessment of the literature on the subject has already been completed. Some scientific papers on contagious diseases in humans have been published, as well as an investigation into how to cure a virus. A dataset of serious corona-virus infectious disease from five different states in India has been considered for clustering in the Pandemic. Based on the silhouette index and Dunn's index of cluster validity, we've been using this technique to investigate for ways to produce a high-quality cluster. Using its dual index formed cluster, the considered dataset was used to examine the cluster compactness. The compression of the cluster aids in limiting virus transmission and limiting the virus's influence on the community to cluster individuals. In this post, we explain the importance of encouraging researchers in this field, as well as some recommendations for the future infectious disease research.
Keywords: Infectious disease, Data set, Clustering approach, contagious illness, .Corona virus diseases.

I. INTRODUCTION

In the months of November and December 2019, a severe acute respiratory disease caused by a novel coronavirus (COVID-19) began spreading across China and drew international attention. The new coronavirus (COVID-19) epidemic was declared a public health emergency of international concern by the World Health Organization (WHO) on January 30, 2020. This outbreak rapidly spread globally and affected the circulatory tract. SARS-COV-2 causes COVID-19, which is a more serious disease. Their dangerous infections might impact either the upper or lower respiratory system. The SARS-COV-2 coronavirus is a modern method of coronavirus, similar to MERS and SARS. The latest n-COVID-19 outbreak has been dubbed "the new coronavirus" over the world. Even though the virus originated in China, it has spread to numerous countries including Italy, South Korea, and Iran. Dry cough, fever, and tiredness are the most prevalent symptoms, while pains, aches, diarrhoea, headache, and conjunctivitis, or sore throat, are less common. The similarities across India's worst-affected districts are revealed using a cluster analysis.

Breathing difficulties, chest pain, difficulty speaking, and bluish lips or cheeks are all significant indications. [1], SARS-Cov-2 is now causing an outbreak in China, with 45171 confirmed cases of pneumonia reported through February 12, 2020 [2]. It was formerly thought to be a form of pneumonia. Virus transmission can occur by a droplet formed by an infected individual coughing or sneezing, contact with infected animals, public areas, meetings, gatherings, and so on.

Protect yourself from n-COVID-19 by wearing a mask, washing your hands correctly with soap, keeping a safe distance from infected or seeking persons, and heating your food before eating it because the virus cannot survive at high temperatures. To protect yourself from the n-COVID-19 virus, keep a safe distance from ill people, wash your hands frequently, and cover your face with a mask. Also, heat your meals thoroughly so the virus cannot survive at high temperatures.

Thus, you should heat up your food properly before consuming it [3].Author review related concept transmission and pathogenesis of MERS-CoV, SARS- CoV, and also discussed the measure to fight its viruses [4]. Virus transmission person to person has been explained familial settings of the cluster [5]. Prevention any spread of deadly coronavirus and its control by setting of healthcare as well as hospital surfaces [6-7].

---

*Corresponding Author: Dr. Vikas Kumar

The infection spreads to the lungs, causing trouble breathing and death. After breathing in the disease, the ailment is transmitted to the next individual who is sensitive to the illness without producing any negative impacts. The illness has a fourteen-day life expectancy from the day it creates side effects including fever, severe coughing, and predictable sniffling to infect more individuals in the contamination region.

In[8], authors are studies and reveals that SARS, MERS or human coronavirus (HCoV) can persist by inanimate surface up to nine days can be stop the outbreak of it thread. The deadly coronavirus pandemic will continue and more researchers are speeding the trail of vaccines in the last stage but while researchers say ensure by end of the year work of vaccination.

The considered dataset have been used to analyze the cluster compactness utilizing its dual index created cluster investigated. The cluster compactness aids in restricting the transmission of the virus and confining its impact on the community to individuals who belong to the cluster. In the present study of this paper the importance of encouraging researchers in this domain, as well as a few suggestions for future studies in infectious disease.

This paper is prepared and organized as follows. In Section- 1 discussed the brief introduction related to the Covid-19, Section-2 discussed the brief literature survey, Section-3 discussed the brief concept of the clustering approach, Section- 4 explains the experimental results, and Section-5 briefs the conclusions and future scope.

II. REVIEW LITERATURE

The deadly corona virus pandemic growth in entire the world and suggestion of the researcher's common symptoms like fever, lack of the appetite, cough, tiredness, pain , breathing problem among people have suffered the COVID-19. Now suppose you feel worse cold with fever, then suspect the COVID-19 and several people have suffered blood clots in arteries, breathing problem feels not better to advise the doctor immediately and rest for fourteen days at home [3,8].

Clustering of data is an unsupervised learning concept to extraction of data and ideas of this method the same type data put in the one cluster and other data place in various clusters..Dunn's index is the second component to the measure of cluster quality, it depend on the minimum separation and max-intra-cluster distance (compactness). The Dunn's index is minimum separation divided by max intra-cluster distance [9]. In [10], author proposed novel cluster validity, modification of Dunn's score and measure the shortest path. The comparatively studied between hard and c-means of fuzzy clustering by the author in [11].

In [12], authors discussed the measure internal quality of cluster: Cohesion of the cluster is sum of assign the weight every links in a cluster its measure by sum of squared errors (SSE) or WSS (Within the cluster Sum of Square), and sum of weights between the node object in cluster and outside the cluster known as the separation (between the cluster sum of square) of the cluster in a graph based cluster. $SSE = WSS = \sum_i \sum_{a \in C_i} (a - m_i)$, $BSS = \sum |C_i|(m - m_i)^2$, $|C_i|$ =Size of the ith cluster.Average silhouette is the measure the cluster validly; pick exact number of cluster, relative quality of the cluster, compactness and the separation Silhouette value is consists both concepts of cohesion and separation [13-14].

Cluster validity with fuzzy clustering technique is supported by similarities and dissimilarities define the pairwise. The silhouette index is generalized and applied on both crisp and fuzzy approaches [15]. The detecting of separated clusters concept in research paper report proposed by Dunn's in [16], and study the survey on cluster [17]. The objective function value is minimized, and show facilitates the new relationship same function, stopping algorithm and prove this relationship [18]. The variable of confirm cases, cure cases, and death cases from 20 May 2021 to 31 July 2021 in India, five different states Rajasthan, Punjab, Uttar Pradesh, Gujarat, and Maharashtra there were more cases spread in these states.

III. CLUSTERING APPROACH

Details of k-means in [20, 21], several steps as given below:
i. Choose k-prototype centroid at random from a data point
ii. Create the initial partitioning of the cluster by assigning the object to the nearest centroid
iii. Calculate the mean of centroid for all clusters.
    a. Any object to find its distance to all centroids.
    b. The object to assign cluster with nearest centroid.
    c. If the object was re-assigned, recalculate the centroids on the basis of new clusters.
iv. Repeat step above (i-iii) until make clusters.

In [14], the quality of cluster is determined by silhouette; therefore silhouette $s_{(k)}$ is clearly defined as

$$H_{(k)} = [B_{(k)} - A_{(k))}]/\min\{A_{(k)}, B_{(k)}\} \qquad (1)$$

Differences of the kth object to other rest object in the own cluster, neighboring cluster and range of silhouette is $-1 \leq H_{(k)} \geq 1$, case: 1- if $H_{(k)}$ close to 1, cluster clearly distinguish (better quality of cluster) orgood cluster, case:2- if $H_{(k)} = 0$, then sample represent the overlying in cluster or not significant, case: 3- if $H_{(k)}$ close to -1, then sample misclassified or create the cluster in wrong direction

And equation- (1) is simplified to (2) we get

$$H(k) = \begin{cases} 1 - [\frac{A(k)}{B(k)}] & if\, A(k) < A(k) \\ 0 & if\, A(k) = B(k) \\ [\frac{B(k)}{A(k)}] - 1 & if\, A(k) > B(k) \end{cases} \qquad (2)$$

In [10], author Ilc (2012), discuss the evaluation of the internal quality of cluster and identifies sets of cluster compactness. Dunn's index introduced by Dunn's in 1974, measure it index as follows the steps;

a.   Measure max distance are given by

$$\Delta_i = max_{A,B \in C_i} dist(A,B), \qquad (3)$$

b. Measure mean distance of all pairs are given by

$$\Delta_i = \frac{2}{(|C_i|(|C_i|-1))}[\sum_{A,B \in C_i, a \neq b} dist(A,B)] \qquad (4)$$

c. Measure the distance from mean are given by

$$\Delta_i = \frac{\sum_{A \in C_i} dist\,(A, mean\,)}{[|C_i|]}, \quad mean = \frac{\sum_{A \in C_i}}{[|C_i|]} \qquad (5)$$

d. Measure Dunn's index use the following formula are given by

$$U = minimum_{i<i<j \leq m} \partial(C_i, C_j)/[maximum_{1 \leq k \leq m\, \Delta k}] \qquad (6)$$

Where, m is the cluster.

IV.   RESULTS

This study examines a clustering technique that was built and evaluated on NCSS 2020, as well as on the Matlab software. Tables 1 and 2 show the results of clustering for COVID-19 average silhouette index and Dunn's index value at different degrees of clusters for various states in India.

Result: 20 August 2020 to 31 October 2020 India, five different states of India
Variables: Death, Confirm cases
Distance Type: Euclidean
Scale Type: Standard Deviation

Dataset 1: Variable confirm cases and death cases from 20 May 2021 to 31 July 2021 India, five different states of India

Table-1: COVID-19 datasets evaluated Based score of clustering for various five states.

S.N.	States	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1.	Rajasthan	0.6360	0.5802	0.5673	0.5294
2.	Punjab	0.6122	0.5907	0.5590	0.5361
3.	UttarPradesh	0.6567	0.6101	0.5623	0.5178
4.	Gujarat	0.6136	0.6231	0.5723	0.5492
5.	Maharashtra	0.6151	0.5709	0.5457	0.5375

Table-2: Dunn's index of clustering on COVID-19 datasets for various five states.

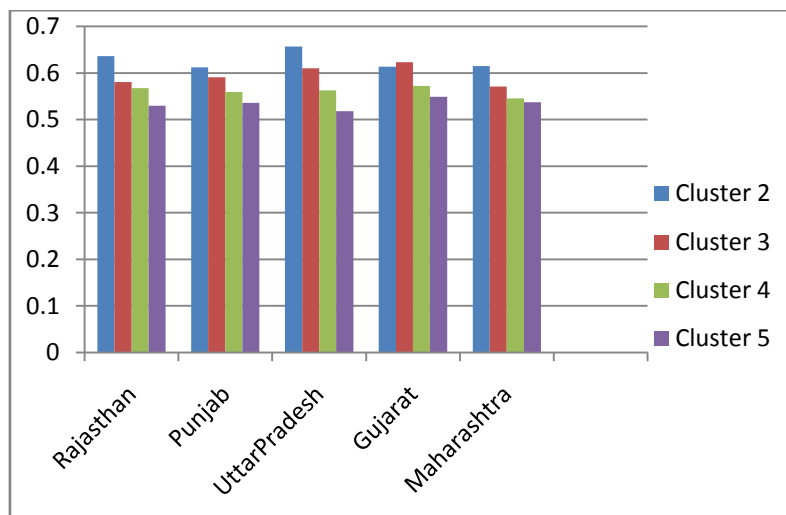| S.N.	States	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1.	Rajasthan	0.7923	0.6209	0.5588	0.5136
2.	Punjab	0.7626	0.6382	0.5756	0.5249
3.	Uttar Pradesh	0.7586	0.6459	0.5764	0.5320
4.	Gujarat	0.7775	0.6604	0.5922	0.5463
5.	Maharashtra	0.7252	0.6132	0.5497	0.5078

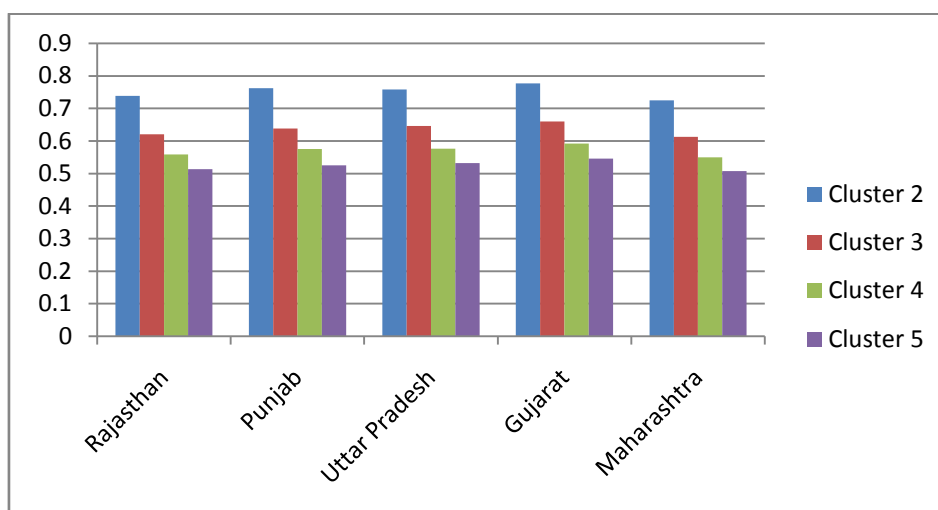Figure 2. Analysis the clusters between silhouette values vs. Clusters

Figure 3. Analysis the clusters between Dunn's Index vs. Clusters.

Dataset 2: Variable confirm cases and cure cases from 20 May 2021 to 31 July 2021 India, five different states of India.

Table-3: Average SilhouetteScore of clustering on COVID-19 datasets for various five states.

S.N.	States	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1.	Rajasthan	0.5623	0.5816	0.5592	0.5319
2.	Punjab	0.6150	0.5923	0.5412	0.5235
3.	Uttar Pradesh	0.6623	0.5236	0.5743	0.5965
4.	Gujarat	0.6815	0.6159	0.5662	0.5345
5.	Maharashtra	0.6235	0.5535	0.5461	0.5072

Table-4: Dunn's index of clustering on COVID-19 datasets for various five states.

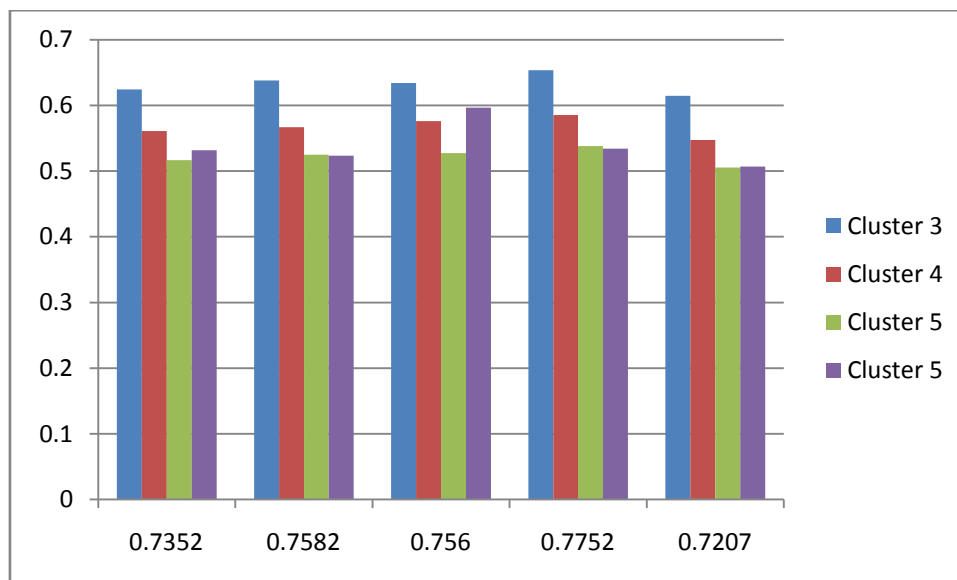| S.N.	States	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1.	Rajasthan	0.7352	0.6243	0.5609	0.5168
2.	Punjab	0.7582	0.6383	0.5671	0.5251
3.	Uttar Pradesh	0.7560	0.6342	0.5763	0.5274
4.	Gujarat	0.7752	0.6537	0.5853	0.5382
5.	Maharashtra	0.7207	0.6145	0.5474	0.5057

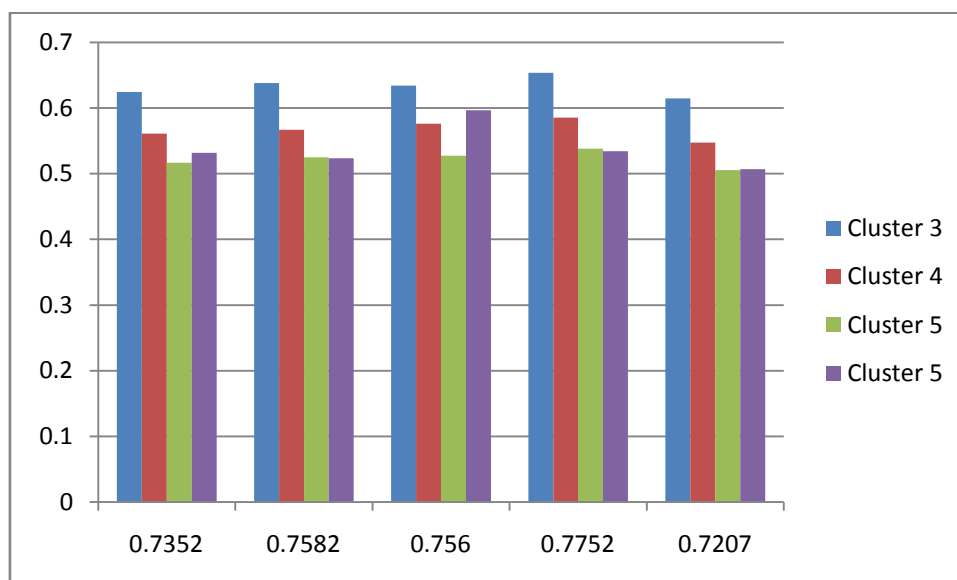Figure 4. Analysis the clusters silhouette values vs. Clusters.

Figure 5. Analysis the clusters on Dunn's index vs. Clusters.

For dataset-1 average Silhouette score of clustering on COVID-19 datasets for various five states in table-1 and Dunn's index of clustering on COVID-19 datasets for various five states in table-2 show the respectively. The analysis of cluster compactness in fig.-2, fig.-3 by silhouette index and Dunn's index respectively compare the five states in India.

The numbers 2, 3, 4, and 5 have been created as a cluster between confirmed cases and deaths. Due to the drop in both silhouette and Dunn's index from 2 to 5, the clusters' compactness is lowered from 2 to 5. tables 3 and 4 show the average Silhouette score of clustering on COVID-19 datasets for various five states, and tables 3 and 4 show the Dunn's index of clustering on COVID-19 datasets for various five states. The silhouette index and Dunn's index were used to compare the five states of India in fig.-4 and fig.-5, respectively. Based on the considered dataset, the created clusters for confirm cases and death cases are 2, 3, 4, and 5.

Major reductions in the index value from 2 to cluster 5, the compactness of clusters is reduced based on both silhouette and Dunn's index from cluster 2 to cluster 5. As a result, the newly created clusters 2, 3, 4, and 5 are well isolated from each other, readily recognizable, and pandemic COVID-19 does not affect their respective clusters.

V.   Conclusion

The correlation between confirmed cases and death cases in India from 20 May 2021 to 31 July 2021, encompassing confirmed instances and recovered cases in the same time frame from five different Indian states, was investigated, as well as the criteria and plans for spreading the revolutionary Covid-19.

The spread of the virus to other parts of India is only getting worse. The numbers of confirmed cases, death cases, and recovered cases range from new corona-virus counts on a daily basis in various Indian states. The Kaggle data sets time series are used to generate the Covid-19 datasets.

The clustering effects suggested that the virus was disseminated similarly in Gujarat, Maharashtra, and Punjab, but not in the other states. The goal of this paper is to examine Covid-19's in Rajasthan, Punjab, Uttar Pradesh, Gujarat, and Maharashtra. We can assess the average silhouette score and Dunn's index by establishing discrete clusters and employing clustering algorithms, albeit compactness is reduced due to the index's lower value and well-created clusters.

The compactness of familiar clusters has been discovered to aid in the prevention and control of infection. Future study should classify artificial intelligence, time series models, and statistical models that can be fitted on the Covid-19 datasets, according to the authors.

References

[1].   Lung Disease :Coronavirus and COVID-19 Overview: Symptoms, Risks, Prevention, Treatment and more. Nov. 28, 2020.
[2].   WHO.: Corona-virus Disease 2019(COVID-19). WHO;2020, Stuation Report 23.
[3].   Patient care.: Coronavirus disease 2019 (COVID-19): Symptoms and causes, Diagnosis and treatment. Nov. 30, 2020.
[4].   De Wit E., Van Doremalen N., Farzarano D., Munster V. J.:SARS and MERS: recent insights into emerging corona-viruses. Nature Reviews Microbiology, 14(8); 523, 2016.
[5].   Chan J. F. W., Yuan S., Kok K. H., To K. K. W., Chu H., Yang J.,..Yuen K. Y.:A familial cluster of pneumonia associated with the 2019 novel corona-virus indicating person to person transmission: astudy of a family cluster. The Lancet, 2020.
[6].   Otter J. A., Donskey C, Yezli S., Douthwaite S., Goldenberg S., Weber D.: Transmission SARS and MERS coronaviruses and influenza virus in healthcare settings: the possible role of dry surface contamination. Journal of Hospital Infection, 2016.
[7].   Dowell S. F., Simmerman J. M., Erdman D. D., Wu J. S. J., Chaovavanich A., Javadi M.,..and, Ho M.: S.Severe acute respiratory syndrome coronavirus on hospital surface.Clinical Infectious Diseases, 2004; 39(5); 652-657.Doi:10.1086/422652
[8].   Kampf G., Todt D., Pfaender S., and Steinmann E.: Persistence of corona-viruses on inanimate surfaces and their inactivation with biocidal agents. Journal of Hospital Infection, 2020; 104; 246-251.
[9].   Bezdek, J. C., and Pal, N. R.: Cluster validation with generalized Dunn's indices. In proceedings 1995 Second New Zealand International Two Stream Conference on Artificial Neural Networks and Expert Systems(pp. 190-193). 1995; IEEE.
[10].   Ilc, N.: Modified Dunn's clustering validity index based on graph theory. Przeglqd Electrotechniczny. 2012; 88(2); 126-131. [Online].Available: http://pe.org.pl/articles/2012/2/36.pdf
[11].   Bora D. J., and Gupta D. A.: A comparative study between fuzzy clustering algorithm and hard clustering algorithm.arXiv preprint arXiv: 1404.6059; 2014.
[12].   Steinbach, M., Kumar, V., and Tan, P.: Cluster analysis: Basic concepts and algorithms. Introduction to data mining,1st edition, Pearson Addison Wesley; 2005.
[13].   Rousseau, P. J.; Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987; 20; 53-65.Doi:10.1016/0377-0427(87)90125-7.
[14].   Silhouette (Clustering): Dec, 28; 2020.
[15].   Rewashes M. and Ralescu A.: Fuzzy cluster validity with generalized silhouette. In annual Meeting of the North American Fuzzy Foundation Processing Society (NAFIPS), 2012;(pp.16).IEEE. [Online]. Available:http://ceur-ws.org/vol-841/submission_33.pdf
[16].   Dunn J. C.: A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. J. Cyberner 3, 1973; .32- 57.
[17].   Ahamad M. K., Bharti A. K.: An Effective Technique on Clustering in Perspective of Huge Dataset. International Journal of Recent Technology and Engineering, 2020; 8(6); 4485-4491.
[18].   Selim S. Z. and Kamel M. S.: On the mathematical and numerical properties of the fuzzy c-means algorithm. Fuzzy sets and System, 1992; 49(2); 181-191.
[19].   Kaufman L.andRousseau P.:FindingGroupsin Data:AnIntroduction to Cluster Analysis.Wiley; 1990.
[20].   Han j., Pei J., Kamber M.: Data mining: concepts and techniques. Elsevier; 2011.
[21].   Anil Kumar, D. Seth, P. Singla and H. K. Gupta: Performance and analysis of Covid-19 in India for spreading -a Mathematical approach, European Journal of Molecular & Clinical Medicine, ISSN 2515-8260 vol.7 (8) , pp 1925-1930, 2020.