Analyzing COVID-19 Virus Behavior by Using K-mediods Clustering Algorithm

Ebtehal Talib [1]
talibebtehal@gmail.com

Abstract: Covid-19 is a virus sweeping all countries of the world like a tornado, but in varying proportions, depending on the health awareness of the country's population as well as the services and health care provided by the country. Despite the great scientific development in the medical field, there is no effective treatment for the virus or a vaccine that has proven to be highly effective. Therefore, countries need non-medical or clinical solutions to confront the epidemic and control its spread by understanding the behavior of a disease through the use of data mining tools. In this research paper, a k-mediod clustering algorithm was applied to a data set of Covid patients in the Philippines, and the results showed the behavior of a disease by collect data according to the age groups that Covid-19 virus targeted it, and the areas in which the virus spread more than others, as well as cases of death, recovery and cases of infection without symptoms, also, severe and mild cases of infection. All of these results clearly reflect the behavior of the virus to provide a complete scenario for the health authorities so that they can take the necessary procedures.

Keywords: Covied-19, Clustering, k-mediods algorithm.

1. Introduction

COVID-19 is considered a dangerous virus because a vaccine and medical treatment have not been proven to be highly effective so far. Despite this, people infected with this virus recover by using some viral medicines, antibiotics, and vitamins Supplements [2]. The world needs a quick solution with the help of non-clinical methods to control and treat the spread of the virus, such as data mining techniques, and the patient database in the Philippines will be used. Clustering is one of the methods for analyzing and discovering the behavior of data, where a subset of elements is selected from data set so that the data elements within a sub-set are more similar to each other while they are more different from the elements of other groups [3]. The similarity between data items are measured using specific measures such as Euclidean distance or correlation scale, Where clustering are formed by comparing the features of the elements using a distance scale [5].

[1] Assist. Lecturer, Ministry of Higher Education and Scientific Research

2. Literature reviews

2.1. Yoon-Jung Choia and eat al [6]:

Dataset	South Korea patients
Methods	clustering using K-means, clusters characterized by size and duration, with >5 patients.
Results	Depending on social distancing, 4,033 patients were classified into three clusters: small, medium and large. The greater social distancing, increased the shift from the large cluster to the small cluster.
Conclusions	By studying the behavior of the virus, the authors were recommend the implementation of more effective strategies in confronting the virus.

2.2. Md. Zubair and eat al [2]:

Dataset	the national strategic data of the COVID-19 epidemic in some nations.
Methods	An improved K-means algorithm has been proposed by the authors by efficiently identifying cluster centers.
Results	Understand disease Prevalence behavior of the disease by country.
Conclusions	The results proved that the improved method was better than the traditional method, by reducing the time required for analysis COVID-19 data patients.

2.3. Shashank Reddy Vadyala1 and eat al [3] :

Dataset	Louisana state USA patients
Methods	A prediction model was constructed using long short-term memory (LSTM) neural networks and K-Means.
Results	extreme gradient boosting was used with weighted k-means algorithm to find similarities between data of past days and forecasts.
Conclusions	The results proved that, the improved algorithm has a higher accuracy compared with the traditional algorithm.

2.4. Yong Shuai and eat al [5]:

Dataset	The national strategic data of the COVID-19 epidemic in some nations.
Methods	A hybrid-clustering model was established based on variety of clustering algorithms (Agglomerative, K-Means, Density-Based with Noise (DBSCAN)).
Results	Studying the results of virus spread based on impact of national epidemiological policies.
Conclusions	The proposed model was proved its accuracy and feasibility.

3. Dataset

The dataset of COVID-19 patients of philipen was obtained from KCDC which was made available on "Kagge Website". The dataset has 5000 instances with 5 attributes which include (case ID, region, age groub, sex,health status) as show in table(1), while table(2) shows a sample of the data set. The data set was divided into two parts: a training set (80%) and a test set (20%).

Table 1:dataset attribute type

Attribute	Descripe Attribute data
Case_id	Integer number (1-5000)
Sex	Male Female
Age-group	-9 10 to 19 20 to 24 25 to 29 30 to 34 35 to 39 40 to 44 45 to 49 50 to 54 55 to 59 60 to 64 65 to 69 70 to 74 75 to 79 80+

Health state	Died Recovered Mild Asymptomatic Critical Severe
Region	Central Visayas (Region VII) Metropolitan Manila CALABARZON (Region IV-A) Central Luzon (Region III) Cordillera Administrative Region (CAR) Davao Region (Region XI) Cagayan Valley (Region II) Western Visayas (Region VI) Ilocos Region (Region I) Bicol Region (Region V) SOCCSKSARGEN (Region XII) Autonomous Region of Muslim Mindanao (ARMM) MIMAROPA (Region IV-B) Northern Mindanao (Region X) Eastern Visayas (Region VIII) Zamboanga Peninsula (Region IX) Caraga (Region XIII) NA

4. Data Mining Technique (Kmediods Algorithm)

There are many clustering method, in this paper used k-medoid because it is more bobust to noisy data.This algorithm used object representative instead of using mean of the objects, as shown in algorithm (1) [1], The algorithm was programmed to obtain the results using the Python language.

```
Algorithm (1): The k mediod  Algorithm
Input:        D = {d1, d2,......,dn}          #set of n vectors with several attributes
              K: Number of clusters
Output:       A:  set of k clusters.
Step 1:  Select k objects from dataset D randomly as an intimal representative;
Step2:   Repeat For all the objects in D
              assign object i to the cluster C
              Randomly select a non-medoid data item
              compute the total cost of swapping old medoid data item with the currently selected non-medoid data item.
              If the total cost of swapping is less than zero, then
              perform the swap operation to generate the new set of k-medoids.
         Until convergence criteria is met
```

5. Model Performance

The performance of clustering algorithm can be determined by measured its accuracy (effectiveness). There are several metrics for measuring the effectiveness of model (mean square error, sum square error, precision, recall,ect). In this paper the sum square error was used. Sum square error objective is to find a clustering that minimizes error using the following equation:

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{x \in c_i} ||x - m_i||^2 \qquad ....(1)$$

Where k refers to number of cluster, m_i is center of cluster [4] . When clustering data set, the sum squared errors is: 5595.656.

Table 2: Sample data of COVID-19 dataset

case_id	Age group	sex	Health status	region
C130591	55 to 59	Female	Died	CALABARZON (Region IV-A)
C178743	35 to 39	Male	Recovered	Metropolitan Manila
C325527	55 to 59	Female	Recovered	Central Luzon (Region III)
C257810	80+	Female	Died	CALABARZON (Region IV-A)
C348794	70 to 74	Male	Recovered	Metropolitan Manila
C473368	20 to 24	Female	Recovered	Davao Region (Region XI)
C537619	75 to 79	Male	Died	CALABARZON (Region IV-A)
C607655	70 to 74	Female	Mild	Central Luzon (Region III)
C615606	70 to 74	Male	Recovered	Metropolitan Manila
C615726	65 to 69	Male	Recovered	Western Visayas (Region VI)
C102418	25 to 29	Female	Mild	Metropolitan Manila
C467667	-9	Female	Recovered	Metropolitan Manila

6. Results and Data analysis

Data were clustered in several ways for the purpose of understanding virus behavior more accurately, as follows:

1. The data were grouped into two clusters according to the **sex** of the person (male- female) infected with the (covied19). The results showed that the

percentage (55%) of those infected with the virus were males, while the percentage of females (45%), as show in figure (1).

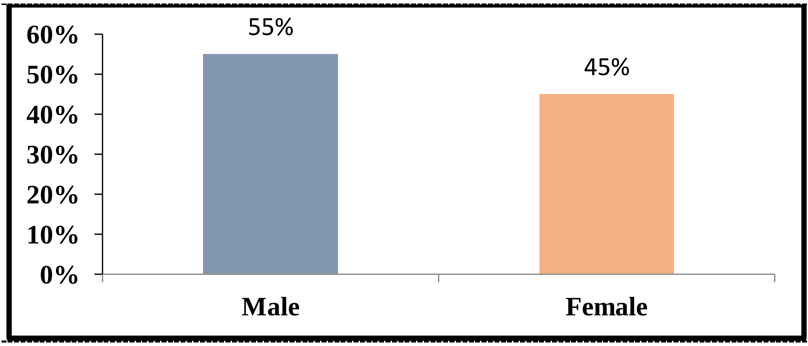


Figure 1. figure (1): Clustering data set according to sex

2. The data were grouped into 15 clusters according to the **Age groupe** of the patient infected with the (covied19). The results showed that the high percentage of the infected in middle age groupes (30 - 70), while the low percentage in small (20-) and old (75+) age groupes, as show in figure (2).

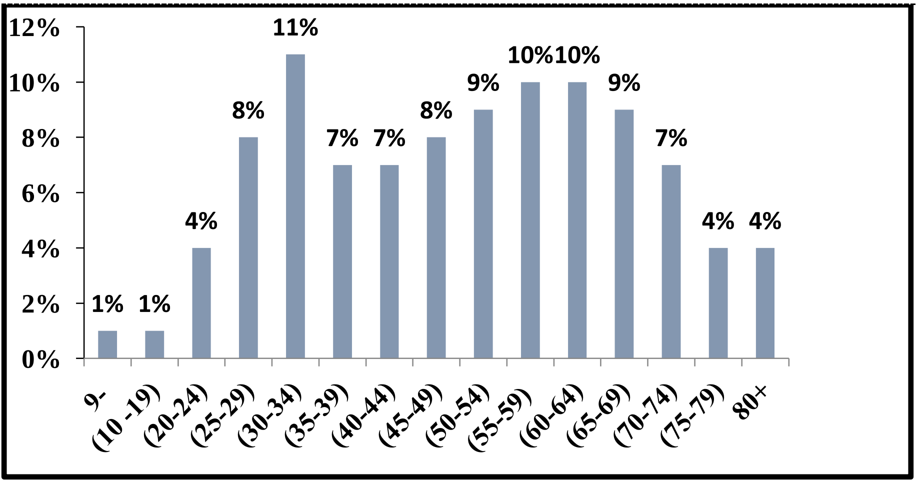
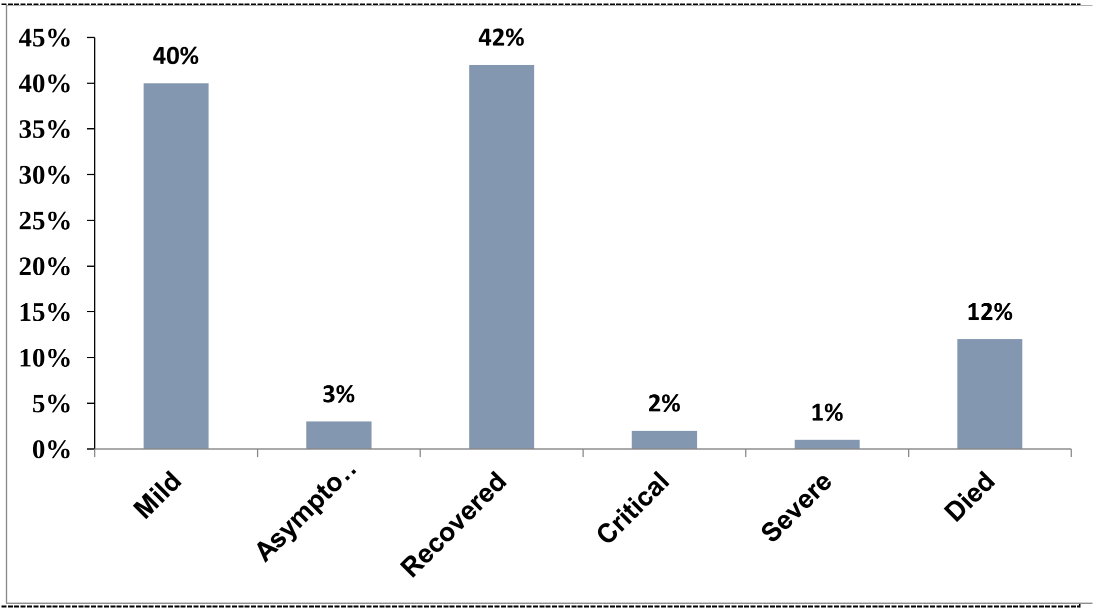


Figure 2. Clustering data set according to Age Groups

3. The data were grouped into 6 clusters according to the **health_status** of the patient infected with the (covied19). The results showed that the percentage of the recovered is (42%), while the percentage of the died is (12%). While the other cases were distributed between Mild by (4%), Asymptomatic by (3%), Critical by (2%) and Severe by (1%), as show in figure (3).



Figure(3): Clustering data set according to health_status

4. The data were grouped into 18 clusters according to The area in which the patient infected with the (covied19) lives. The results showed. The virus-endemic area is (Metropolitan Manila), while the spread of the virus was slight in the rest of the areas, as show in figure (4).

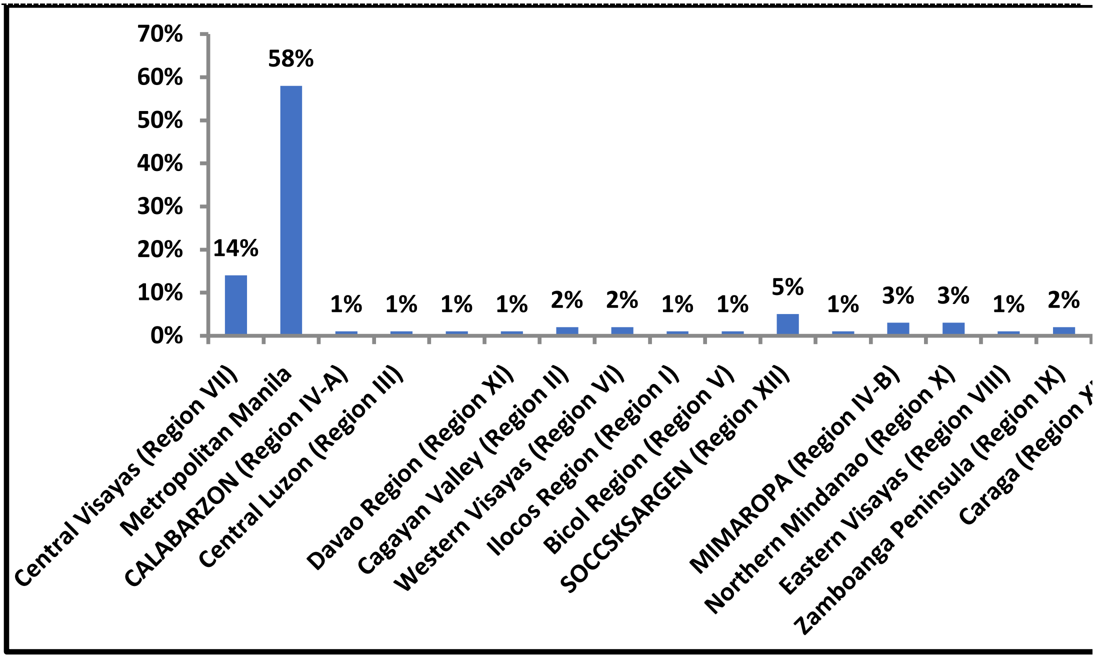


Figure (4): Clustering data set according to Regions

7. Conclusions

This paper aims to aid specialists to understand the spread of COVID-19 in Philippine. Through the available dataset studies, The areas in which the disease spread severely and the areas where the number of infections were few were identified, and the age groups most targeted by the disease were identified, as well as the diversity in the severity of disease cases, the number of recovery and dead cases. so this paper contributes to enabling the health authorities to take appropriate measures to prevent virus Spread to rest of the regions, Taking more severe preventive measures in endemic areas and spreading health awareness more among different age groups, As well as health authorities became more knowledge when facing the next waves of corona virus life-threatening.

8. References

[1] Mahendra Tiwari and Randhir Singh (2012) "Comparative Investigation of K-Means and K-Medoid Algorithm on Iris Data" *International Journal of Engineering Research and Development, Volume 4, Issue 8*.

[2] Md. Zubair, MD.Asif Iqbal, Avijeet Shil1, Enamul Haque, Mohammed Moshiul Hoque and Iqbal H. Sarker(2020) *"An Efficient K-means Clustering Algorithm for Analysing COVID-19"*, arXiv:2101.03140v1.

[3] Shashank Reddy Vadyala, Sai Nethra Betgeri Eric A. Sherer Amod Amritphale (2020) *” Prediction of the Number of COVID-19 Confirmed Cases Based on K-Means-LSTM”,* https://arxiv.org/abs/2006.14752.

[4] Yedla ,Madhu, Pathakota ,Srinivasa Rao and Srinivasa, T. M. (2010) *“Enhancing K-means Clustering Algorithm with Improved Initial Center”* International Journal of Computer Science and Information Technologies, Vol. 1 (2) ,pp. 121-125.

[5] Yong Shuai, Chunxu Jiang, Xinyi Su, Can Yuan and Xiaoping Huang(2020) *” A Hybrid Clustering Model for Analyzing COVID-19 National Prevention and Control Strategy “*,IEEE 6th International Conference on Control Science and Systems Engineering, Chongqing, China.

[6] Yoon-Jung Choia and eat al (2020) *”Type of COIVED19 clusters and theire relationship with social distancing in the Seoul metropolitan in south Korea”,* International journal of infectious Diseases, 202.

تحليل سلوك فيروس COVID-19 باستخدام خوارزمية (mediods) للتجميع

ابتهال طالب خضير[1]
talibebtehal@gmail.com

المستخلص: Covid-19 هو فيروس يجتاح جميع دول العالم مثل الإعصار، ولكن بنسب متفاوتة اعتمادًا على الوعي الصحي لسكان البلد بالإضافة إلى الخدمات والرعاية الصحية التي تقدمها الدولة. على الرغم من التطور العلمي الكبير في المجال الطبي، الا انه لا يوجد علاج فعال للفيروس أو لقاح أثبت فعاليته العالية. لذلك، تحتاج الدول إلى حلول غير طبية أو سريرية لمواجهة الوباء والسيطرة على انتشاره من خلال فهم سلوك المرض من خلال استخدام أدوات التنقيب عن البيانات. في هذه الورقة البحثية، تم تطبيق خوارزمية التجميع k-mediods على مجموعة بيانات لمرضى كوفيد في الفلبين، وأظهرت النتائج نهج سلوك المرض من خلال تجميع البيانات وفقًا للفئات العمرية التي استهدفها فيروس كوفيد -19، و المناطق التي ينتشر فيها الفيروس أكثر من غيرها، وكذلك حالات الوفاة والشفاء وحالات الإصابة بدون أعراض، وكذلك حالات العدوى الشديدة والخفيفة. كل هذه النتائج تعكس بوضوح سلوك الفيروس لتقديم سيناريو كامل للجهات الصحية حتى تتمكن من اتخاذ الإجراءات اللازمة.

الكلمات المفتاحية : كوفييد -19، التجميع، خوارزمية k-mediods

[1] مدرس مساعد؛ وزارة التعليم العالي و البحث العلمي