Mapping the Spread of Covid-19 in Asia Using Data Mining X-Means Algorithms

N Arminarahmah, A Daengs GS, G W Bhawika, M P Dewi and A Wanto

1 Universitas Islam Kalimantan Muhammad Arsyad Al Banjari, Banjarmasin, Indonesia
2 Universitas 45 Surabaya, Surabaya, Indonesia
3 Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
4 Institut Teknologi dan Bisnis Asia Malang, Malang, Indonesia
5 STIKOM Tunas Bangsa, Medan, Indonesia

E-mail: anjarwanto@amiktunasbangsa.ac.id

Abstract. Until now, Covid-19 is a phenomenal problem faced by almost all over the world, especially countries on the Asian continent. Not only causing casualties, but this virus also affects the wheels of the country's economy. The purpose of this paper is to view and map the spread of the Covid-19 virus in Asia based on Total Cases, Total Deaths, Total Cured and Active Cases from 49 Countries. The research data in this paper were obtained from the Worldometer website sourced from WHO, CDC, NHC and others. In this proposed paper, the algorithm used is X-Means Clustering with the help of Rapidminer. The results of this proposed paper are in the form of grouping or mapping the spread of the Covid-19 virus in Asia which is divided into 4 zones, including the Red Zone (the number of active cases of Covid-19 and the death rate is very high) which consists of 1 country, the Orange Zone (number of cases active covid-19 and the mortality rate is quite high) consisting of 1 country, the Yellow Zone (active case rate of Covid-19 and moderate mortality) consisting of 39 countries, and the Blue Zone (active case rate of covid-19 and low mortality rate) which consists of 8 countries.

1. Introduction

The coronavirus pandemic until the end of 2020 is still a phenomenal problem faced by almost all over the world since it was first identified in Wuhan last December 2019 and it is not certain exactly when it will end. This includes countries in the Asian continent, which is the largest continent in the world. As previously known, the COVID-19 disease, which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a major global crisis [1][2]. On March 11, 2020, the World Health Organization (WHO) declared COVID-19 a pandemic, and it has become one of the deadliest pandemics in the past century [3]. Not only causing casualties, but this virus also affects the wheels of the country's economy. This paper will discuss the mapping of the distribution of Covid-19 in Asian countries, which consist of 49 countries, this is because this virus was first identified in China, which is one of the countries on the Asian continent. Based on data obtained from the Worldometers website sourced from WHO, CDC and NHC, the number of Covid-19 cases in the world as of Sunday 13 December 2020 at 11.22 WIB reached 72,653,057 people with a total death reaching 1,618,890
people, while in Asia from In 49 countries, the number of Covid-19 cases reached 19,084,461 people with a total death toll of 312,236 [4].

Therefore, mapping the distribution of Covid-19 is very important to do, as information and a barometer for governments in Asian countries, especially for the Indonesian government in determining policies related to this matter. The specific purpose of this research is to provide input and information for the Indonesian government to maximize efforts to be able to overcome the Covid-19 problem and for the Indonesian people to temporarily not travel to countries in Asia that are included in the Red, Orange and Yellow Zones. The research dataset is a dataset of Covid-19 cases in Asia from the Worldometers website sourced from WHO, CDC and NHC. The mapping algorithm used in this study is the X-Means Algorithm.

Table 1. Data on Covid-19 Cases in Asia

| No | Country      | Total Cases | Total Deaths | Total Recovered | Active Cases |
|----|--------------|-------------|--------------|----------------|--------------|
| 1  | Afghanistan  | 49.273      | 1.971        | 38.384         | 8.918        |
| 2  | Armenia      | 148.325     | 2.486        | 125.169        | 20.670       |
| 3  | Azerbaijan   | 175.874     | 1.922        | 111.918        | 62.034       |
| 4  | Bahrain      | 89.143      | 348          | 87.182         | 1.613        |
| 5  | Bangladesh   | 490.533     | 7.052        | 420.896        | 62.585       |
| 6  | Bhutan       | 437         | 0            | 400            | 37           |
| 7  | Brunei       | 152         | 3            | 147            | 2            |
| 8  | Cambodia     | 359         | 0            | 307            | 52           |
| 9  | China        | 86.725      | 4.634        | 81.785         | 306          |
| 10 | Cyprus       | 15.101      | 78           | 2.057          | 12.966       |
| 11 | Georgia      | 189.726     | 1.790        | 158.183        | 29.753       |
| 12 | Hong Kong    | 7.542       | 117          | 6.202          | 1.223        |
| 13 | India        | 9,884.716   | 143.393      | 9,387.608      | 353.715      |
| 14 | Indonesia    | 617.820     | 18.819       | 505.836        | 93.165       |
| 15 | Iran         | 1,108.269   | 52.196       | 812.270        | 243.803      |
| 16 | Iraq         | 574.634     | 12.579       | 507.446        | 45.190       |
| 17 | Israel       | 357.176     | 2.999        | 337.186        | 16.991       |
| 18 | Japan        | 177.287     | 1.790        | 158.183        | 19.790       |
| 19 | Jordan       | 259.614     | 3.365        | 218.467        | 41.147       |
| 20 | Kazakhstan   | 141.578     | 2.088        | 125.380        | 16.198       |
| 21 | Kuwait       | 146.218     | 117          | 139.204        | 6.996        |
| 22 | Kyrgyzstan   | 77.356      | 1.307        | 70.192         | 5.857        |
| 23 | Laos         | 41          | 0            | 33             | 8            |
| 24 | Lebanon      | 146.520     | 1.200        | 101.334        | 43.896       |
| 25 | Macao        | 46          | 0            | 46             | 0            |
| 26 | Malaysia     | 83.475      | 415          | 69.393         | 13.667       |
| 27 | Maldives     | 13.368      | 48           | 12.706         | 614          |
| 28 | Mongolia     | 907         | 0            | 384            | 523          |
| 29 | Myanmar      | 108.342     | 2.687        | 86.795         | 19.279       |
| 30 | Nepal        | 248.423     | 1.698        | 235.731        | 10.994       |
| 31 | Oman         | 126.240     | 1.471        | 118.048        | 6.721        |
| 32 | Pakistan     | 438.425     | 8.796        | 383.000        | 46.629       |
| 33 | Palestine    | 109.738     | 978          | 85.110         | 23.650       |
| 34 | Philippines  | 449.394     | 8.733        | 418.682        | 21.979       |
| 35 | Qatar        | 140.961     | 240          | 138.490        | 2.231        |
| 36 | S. Korea     | 42.766      | 580          | 31.814         | 10.372       |
| 37 | Saudi Arabia | 359.888     | 6.048        | 350.549        | 3.291        |
| 38 | Singapore    | 58.320      | 29           | 58.208         | 83           |
| 39 | Sri Lanka    | 32.790      | 152          | 23.793         | 8.845        |
| 40 | Syria        | 9.166       | 518          | 4.376          | 4.272        |
| 41 | Taiwan       | 736         | 7            | 606            | 123          |
| No | Country          | Total Cases | Total Deaths | Total Recovered | Active Cases |
|----|-----------------|-------------|--------------|-----------------|--------------|
| 42 | Tajikistan      | 12,704      | 88           | 12,133          | 483          |
| 43 | Thailand        | 4,209       | 60           | 3,923           | 226          |
| 44 | Timor-Leste     | 31          | 0            | 30              | 1            |
| 45 | Turkey          | 1,836,728   | 16,417       | 1,603,780       | 216,531      |
| 46 | UAE             | 184,949     | 617          | 163,679         | 20,653       |
| 47 | Uzbekistan      | 74,956      | 612          | 72,243          | 2,101        |
| 48 | Vietnam         | 1,397       | 35           | 1,241           | 121          |
| 49 | Yemen           | 2,083       | 606          | 1,383           | 94           |
|    | Total           | 19,084,461  | 312,236      | 17,265,905      | 1,506,320    |

X-Means is a data mining algorithm which is a development of the K-Means, and this algorithm can cover the shortcomings of the K-Means [5]. As is well known, data mining algorithms are algorithms that are often used for data classification [6]–[10], data clustering [11]–[14] and Prediction [15]–[18]. Several previous studies related to the topic of this article include: Geo-Social Big Data Clustering using Focused Information Criterion based Partitioned Iterative X-Means Dice Correlation Data Clustering (FIC-PIXDCDC) Method. In this algorithm, geosocial data taken as input (user, location, and time) is obtained from the Weeplaces dataset. After that, this algorithm selects the number of clusters and centroids randomly. Then calculate the Dice Correlation between each geosocial data input and the cluster centroid. Furthermore, this algorithm applies Focused Information Criteria to build the optimal number of clusters for a certain large dataset. The results of this study indicate that this algorithm has higher accuracy and shorter time than other conventional algorithms [19]. Furthermore, the use of the Euclidean Distance X-Means algorithm for the BigData cluster. This algorithm is used to test data based on the centroid value which is then modified to obtain five clusters, based on the Euclidean distance. After testing, the data obtained from the cluster members have a similarity level to other data. The number of clusters determined by modification of the euclidean distance resulted in a better similarity level than if it was done randomly from each member [20].

The next research was to determine the X-Means centroid clustering using the Davies-Bouldin method. In this paper the X-Means algorithm is modified to perform several centroid determinations. So as to produce a better level of similarity of cluster members compared to other data. The Davies-Bouldin Index method is used to determine the centroid for testing two clusters with a minimum value with a DBI value close to 0 [21]. These related studies are behind the conduct of research to map the distribution of Covid-19 in Asia.

2. Methodology

2.1. Algorithms and Research Data
This study uses the X-Means Algorithm which is one of the Data Mining algorithms used to cluster data, while the research data is data on Covid-19 Cases in Asia obtained from the Worldometers website sourced from WHO, CDC and NHC (can be seen in Table 1).

2.2. X-Means Stages
The X-Means clustering stages in this study are presented in Figure 1.
Defined $K_{\text{min}}$ and $K_{\text{max}}$ as the lower and upper limits for possible $K$ value $K$. In the initial X-Means clustering step, it is known that currently $K = K_{\text{min}}$. Initial and structures centroid are found by K-Means. In the next stage, each cluster in the presumed structure is treated as the parent cluster, which can be broken down into two child clusters. Based on several criteria, which will be described in the next section, we assessed the structure of the parent and child. Scores help to decide whether parents are a better representation of the sample data or clusters of children provide a more accurate distribution of the sample. As a result, the parent centroid will be replaced by the child centroid, or the algorithm will retain the parent centroid and leave the children. Then, new structures will be built or updated based on the choice of the parent or child. This procedure will continue for all clusters in the initial structure until the current estimated number of clusters is greater than $K_{\text{max}}$ or the algorithm converges to the best structure.

3. Results and Discussion

3.1. Cluster Determination ($K_{\text{min}}$ and $K_{\text{max}}$)

In implementing the X-Means algorithm using Rapidminer, first is to determine the $K_{\text{min}}$ value (the minimum number of clusters that must be detected) and the $K_{\text{max}}$ value (the maximum number of clusters that must be detected). The value of $K_{\text{min}}$ and $K_{\text{max}}$ can be determined randomly or randomly which is optional. There are 4 clusters defined in this paper.

3.2. Process X-Means with Rapidminer

The process of mapping the distribution of Covid-19 in Asia in the form of clustering and the results of the X-Means algorithm using Rapidminer can be seen in figure 2, figure 3, figure 4, figure 5 and figure 6.
Figure 2 describes the X-Means clustering process with Rapidminer which begins by entering Covid-19 data in Asia, followed by the selection of the X-Means operator with a value of $K_{\text{min}} = 4$ and a value of $K_{\text{max}} = 100$. Then it is connected to the Apply Model to apply a model that has been trained or studied. Then connect to Cluster Distance Performance to evaluate the performance of the model which provides a list of performance criteria values automatically according to the given task.

![Cluster Model]

Figure 3. Cluster (X-Means)

Figure 3 is a cluster model produced by Rapidminer, Cluster_0: 39 items, Cluster_1: 1 item, Cluster_2: 1 item, and Cluster_3: 8 items. For the final result, the Centroid table can be seen in Figure 4.

![Centroid Table]

Figure 4. Centroid Table

The detailed results of mapping the distribution of Covid-19 in Asia can be seen in Figure 5.

![Folder View Covid-19 Mapping and Spread in Asia]

Figure 5. Folder View Covid-19 Mapping and Spread in Asia
Cluster_0 based on figure 5 consists of 39 countries (Afghanistan to Yemen). Cluster_1 consists of 1 Country (India). Cluster_2 consists of 1 country (Turkey). Cluster_3 consists of 8 countries (Bangladesh, Indonesia, Iran, Iraq, Israel, Pakistan, Philippines and Saudi Arabia).

The blue dot in the figure is the Covid-19 Cluster_0 (Blue Zone) Spread Mapping. The green dot represents the Covid-19 Cluster_3 Spread Mapping (Yellow Zone). The orange dot represents the Covid-19 Cluster_1 (Red Zone) Spread Mapping. The black dot represents the Covid-19 Cluster_2 Spread Mapping (Orange Zone).

4. Conclusion
Data Mining X-Means can be used to map the Covid-19 spread in Asia. Based on the results of the Covid-19 spread mapping analysis using Rapidminer, it is divided into 4 zones, namely the Red Zone (the number of active cases of Covid-19 and the death rate is very high) consisting of 1 country, namely India (Cluster_1), the Orange Zone (Number of active cases of Covid-19 and the mortality rate is quite high) consists of 1 country, namely Turkey (Cluster_2), the Yellow Zone (the number of active cases of Covid-19 and moderate deaths) consists of 8 countries, including: Bangladesh, Indonesia, Iran, Iraq, Israel, Pakistan, Philippines and Saudi Arabia (Cluster_3), and the Blue Zone (the number of active cases of Covid-19 and the low mortality rate) which consists of 39 countries, including Afghanistan to Yemen (Cluster_0), as presented in figure 5. Performance Vector X-Means in this paper results in the value Avg._within_centroid_distance = -20930456418.490 and Davies_Bouldin = -0.266.

References
[1] Z. J. Cheng and J. Shan, “2019 Novel coronavirus: where we are and what we know,” Infection, vol. 48, no. 2, pp. 155–163, 2020.
[2] F. He, Y. Deng, and W. Li, “Coronavirus disease 2019: What we know?,” Journal of Medical Virology, vol. 92, no. 7, pp. 719–725, 2020.
[3] M. R. Azarpazhooh et al., “COVID-19 pandemic and burden of non-communicable diseases: An ecological study on data of 185 countries,” Journal of Stroke and Cerebrovascular Diseases, vol. 29, no. 9, pp. 1–9, 2020.
[4] Worldometer, “Reported Cases and Deaths by Country, Territory, or Conveyance,” COVID-19 Coronavirus Pandemic, 2020. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 15-Dec-2020].
[5] A. Radwan et al., “X-means Clustering for Wireless Sensor Networks,” *Journal of Robotics, Networking and Artificial Life*, vol. 7, no. 2, pp. 111–115, 2020.

[6] I. S. Damanik, A. P. Windarto, A. Wanto, Poningsih, S. R. Andani, and W. Saputra, “Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm,” *Journal of Physics: Conference Series*, vol. 1255, no. 1, pp. 1–6, Aug. 2019.

[7] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto, and A. Wanto, “C.45 Classification Rules Model for Determining Students Level of Understanding of the Subject,” *Journal of Physics: Conference Series*, vol. 1255, no. 1, pp. 1–7, 2019.

[8] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. Perdana Windarto, “Application of Classification Method C4.5 on Selection of Exemplary Teachers,” *Journal of Physics: Conference Series*, vol. 1235, no. 1, 2019.

[9] I. Parlina et al., “Naive Bayes Algorithm Analysis to Determine the Percentage Level of visitors the Most Dominant Zoo Visit by Age Category,” *Journal of Physics: Conference Series*, vol. 1255, no. 1, pp. 1–5, 2019.

[10] D. Hartama, A. Perdana Windarto, and A. Wanto, “The Application of Data Mining in Determining Patterns of Interest of High School Graduates,” *Journal of Physics: Conference Series*, vol. 1339, no. 1, pp. 1–7, 2019.

[11] M. A. Hanafiah, A. Wanto, and P. B. Indonesia, “Implementation of Data Mining Algorithms for Grouping Poverty Lines by District/City in North Sumatra,” *International Journal of Information System & Technology*, vol. 3, no. 2, pp. 315–322, 2020.

[12] N. A. Febriyati, A. D. GS, and A. Wanto, “GRDP Growth Rate Clustering in Surabaya City uses the K- Means Algorithm,” *International Journal of Information System & Technology*, vol. 3, no. 2, pp. 276–283, 2020.

[13] S. Sudirman, A. P. Windarto, and A. Wanto, “Data Mining Tools | RapidMiner: K-Means Method on Clustering of Rice Crops by Province as Efforts to Stabilize Food Crops In Indonesia,” *IOP Conference Series: Materials Science and Engineering*, vol. 420, no. 012089, pp. 1–8, 2018.

[14] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif, and S. Z. H. Kazmi, “K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir,” *IEEE Access*, vol. 8, pp. 151847–151855, 2020.

[15] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, “Implementing AutoML in Educational Data Mining for Prediction Tasks,” *Applied Sciences*, vol. 10, no. 90, pp. 1–27, 2020.

[16] A. K. Verma, S. Pal, and S. Kumar, “Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method-a Comparative Study,” *Applied Biochemistry and Biotechnology*, pp. 1–19, 2019.

[17] S. Borah, V. E. Balas, and Z. Polkowski, *Advances in Data Science and Management*. Springer, 2019.

[18] S. Q. Salih et al., “River suspended sediment load prediction based on river discharge information: application of newly developed data mining models,” *Hydrological Sciences Journal*, vol. 65, no. 4, pp. 624–637, 2020.

[19] M. Anoop and P. Sripriya, “Focused information criterion based partitioned iterative X-means dice correlation clustering for big Geo-social data,” *Journal of Critical Reviews*, vol. 7, no. 6, pp. 54–62, 2020.

[20] N. Zendrato, H. W. Dhany, N. A. Siagian, and F. Izhari, “Bigdata Clustering using X-means method with Euclidean Distance,” *Journal of Physics: Conference Series*, vol. 1566, no. 1, p. 012103, 2020.

[21] M. Mughnyanti, S. Efendi, and M. Zarlis, “Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation,” *IOP Conference Series: Materials Science and Engineering*, vol. 725, no. 1, p. 012128, 2020.