Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.47750/jptcp.2023.30.04.028

Big Data Analytics in Healthcare: COVID-19 Indonesia Clustering

Johanes Fernandes Andry[1*], Glisina Dwinoor Rembulan[2], Edwin Leonard Salim[3], Endang Fatmawati[4], Hendy Tannady[5]

[1,3]Department of Information System, Universitas Bunda Mulia, Jakarta, Indonesia
[2]Department of Industrial Engineering, Universitas Bunda Mulia, Jakarta, Indonesia
[4]Information & Public Relations Study Program, Universitas Diponegoro, Indonesia
[5]Department of Management, Universitas Multimedia Nusantara, Indonesia
***Corresponding author:** Johanes Fernandes Andry, Department of Information System, Universitas Bunda Mulia, Jakarta, Indonesia, Email: jandry@bundamulia.ac.id

Submitted: 13 January 2023; Accepted: 20 February 2023; Published: 05 March 2023

ABSTRACT

The rapid growth of the Internet and Technology produced a massive amount of data that resulted a phenomenon called Big Data. To process such a complex kind of massive amount of data, an advanced approach and tool is needed that is able to quickly produce results. This approach to analyzing massive amount of data is known as Big Data Analytics. Big data analytics is widely used in various sectors, not to mention the health sector. In the healthcare sector, recently there has been a study that is often carried out in dealing with crisis situations, namely research on implementing big data analytics to provide technological solutions to help deal with pandemics. In this article, we analyze and visualize the data collected from Indonesia. The data analyzed starts from the first case of COVID-19 in Indonesia to present. The proposed solution is to classify the regional case data into a group that can represent the situation of the area. As a result, it is determined based on the data that there are three groups consisting of areas with low risk, moderate risk, and high risk. In addition, this article proposes combining big data analytics technology with cloud technology to facilitate the dissemination of information to citizens to increase awareness about the spread of the COVID-19 virus.

Keywords: *Big Data, Big Data Analytics, Data Mining, COVID-19, Clustering, k-means algorithm*

INTRODUCTION

The rapid growth of the Internet and Technology produced a massive amount of data that resulted a phenomenon called Big Data. Big Data generally used to describe an enormous datasets could grow until the amount of data are unmanageable [1][2]. Unmanageable in this context means, the data reside in the storage has grown and keep growing so fast and exceeded the capacity of conventional software that commonly used, which mades those sofware not capable to process data within 'tolerable' elapsed time.[3]. These Big Data phenomenon had 3V core characteristic which is 1) Volume, define the scale of the data, 2) Velocity, define the speed of the data, and 3) Variety, which define the variety of the data. One approach to analyze and leveraging such large, diverse and complex data is known as big data analytics.

Big Data Analytics is a set of action to analyze and examine a complex and large data sets, in this context which is Big Data, and reveal the hidden information contained in the data sets that can help organization or institution with better decision making [4]. As the big data emerged, many institution form varying sector initiate to leveraging this phenomenon to benefit their organization by capturing and utilizing their data, healthcare is no exception. In healthcare, big data analytics known as a large scale technological dataset which very large in size, diverse in variety and high complexity which made those data hard to manage with the traditional method and tools [5]. In healthcare sectors, mainly the research conducted to utilize accumulated data to discover new things which can be useful in this sectors. Furthermore, the usage of BDA on the health sector also helps the executive on the sectors make a better decision based on data they had, in other words the usage of BDA has led the decision taken on healthcare institution is more data-driven [6]. One of the BDA implementation that are popular lately is COVID-19 Big Data Analysis.

Since the end of 2019, the world has been shocked by the spread of a new virus called coronavirus. WHO has designated it a global pandemic and called it COVID-19 [7]. In December 2019, a new form of virus, Corona Virus Disease 2019 (COVID-19) which the first case was registered in Wuhan, Hubei Province, originating from China [8]. Since then the virus started to spread around regions in China. Virus spread around the regions in China by cumulative small incidence that causes the virus infecting more people and the number increasing exponentially until the virus unable to be contained and spreading throughout the world. In Indonesia, the first two cases was announced on March 2020. And as of March 29, COVID-19 the fatalities cases in Indonesia reach the 102 deaths out of 1,155 confirmed cases, which indicates a high risk from COVID-19 in Indonesia. Since March 15, The President of Indonesian, Joko Widodo has advised the citizens to help government reducing the spread of the virus by doing some activities at home for two weeks, to reduce the sprad of coronavirus [9][24].

If the coronavirus pandemic is not handled properly, it will cause many victims and have a major impact on the economy which can result in chaos. In the last few months, many countries has proposed many models to analyze the COVID-19 spread behaviour based on data they had, like in India, many journals proposed a prediction models to predict when COVID-19 ends based on the growth per-day [10], the same like in Nigeria, on the journal [11], author state that Nigeria faces challenges where the test can be performed in a day limited to 60 test per day. So they build a prediction models based on the single source data that are available in their country. Surpsingly, the result the models they proposed show a precise number of the case growth per day. From the small example from that have been done by other countries we can learn and propose a model that could fit with the Indonesia condition [25].

As long the stable vaccine is developed, the only one solution to prevent the virus spreading to infect more people is by doing social distancing in order to minimalize contact. There is another approach that might be helpful, by analyzing the coronavirus big data to gain more clarity and provide a big picture regarding the condition in some country or even smaller region. In Indonesia, due to lack of capability in many aspects, we could not propose a prediction model that could be accurate and took a lot of time and resources to fulfill the requirement as the crisis

J Popul Ther Clin Pharmacol Vol 30(4):e290–e300; 05 March 2023.

emerge day to day pace, and growing exponentially. Therefore, this article proposed a clustering model that might be utilized to groups province based on coronavirus active cases. By grouping the provinces may help government to create a regulation to reduce the coronavirus growth on particular province and contain the pandemic to prevent more outbreak occured. The clustering solution we proposed can be combined with other technologies such as cloud computing to better information dissemination. After the growth per day can be reduced, as the results we may predict the likelihood when the coronavirus ending in Indonesia. Furthermore will be discussed on the "Analysis and Results" section.

The remainder of this paper is divided into 5 sections as follows: "Literature Review" which contains the related literature for the basis theory and previous research of big data analytic conducted on this paper. "Research method" section contains a description of the research methodology used on the paper. "Analysis and Discussion" section contains empirical tests and deep results discussion regarding the model that proposed, and "Conclusion" section contains the conclusion regarding the study and analysis conducted.

2. LITERATURE REVIEW

This section will briefly explain some of the important concepts that underlying this study, including Big Data, Data Mining, Clustering and also some relevant works done by other researchers.

2.1 Relevant Works

Before this study was conducted, there is some other big data analytics implementation with the same scope, which is a study to find out how the COVID-19 behave based on data collected from the each country. The study provided in Table 1: Relevant Works, showing some important aspects that relevant to this study.

TABLE 1: Relevant Works

Authors	Year	Title	Important Aspects
Tuli Et al.[10]	2020	Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing	Leveraging dataset that collected to provide a clear representation regarding the Coronavirus pandemic situation in India Combining machine learning and cloud computing technology to provide value from the data collected
Abdulmajeed, Adeleke, Popoola [11]	2020	Online Forecasting of COVID-19 cases in Nigeria using limited data	Being able to utilize small and limited dataset to predict the likelihood of COVID-19 cases
Mahmudan, Ali [12]	2020	Clustering of District or City in Central Java Based COVID-19 Case Using K-Means Clustering	Applying the clustering approach to group some cities into a clusters to help government

2.2 Big Data

Most of the time, big data known as a big chunk of data, but that does not define big data itself as a whole. Big data is a concept that is abstract. It has other characteristics, apart from masses of data, which specify the distinction between itself and "very big data" or "massive data" [1]. Big Data begins with distributed and decentralized control of large-volume, heterogeneous, autonomous sources and aims to explore dynamic and changing data relationships. [3]. Because of the abstract definition of big data, some of its characteristic must be explored in order to have a better understanding about this concept [26].

There's 3 main factor of data that make the data set called Big Data. The 3 factor is known as 3Vs. First of the factor is Volume which mean the size of data set must be massive. Technological information advancement has resulted so many ways to generate data which led to produce massive chunk of data. As of 2012, on average, about 2.5 exabytes of data are produced per day and that amount doubles every approximately 40

months [13]. As an example, in a day, Google processes around hundreds of Petabytes data, Facebook has a generated logging system which generate over 10 Petabytes per month. From the other side of the world, Baidu, as an Chinese version of Google, processes around tens of Petabytes, and to add more example, Taobao, a Chinese e-commerce company generates a tens of Petabytes [1]. And with all of those big size, the data set still couldn't be called big data.

The second factor is Velocity which mean velocity of growth data. The advancement of information technology made the process of producing data much faster than before. [14]. That's make the data getting bigger, growing exponentially and becoming a big data.

The last factor is Variety which mean variety of format data. Over the time, there is many new formats of data emerged. As a by-product of their daily activities, data generators such as cell phones, internet browsing, social networks, electronic networking, GPS, and designed machines all generate data torrents. [13]. All of this certainly make many of format data like mp3, mp4, JPEG, org GPS signal and many more that make the data very heterogeneous.

2.3 Data Mining

Data mining is a techniques to unveil useful insight and hidden patterns from data by collecting, cleaning, processing, analyzing data sets. In terms of problem domains, application, formulation, and representations, there is a wide variation of data mining In terms of the problem domains, implementations, formulations, and data representations found in actual applications, there is a wide variety. [15]. From other perspective, Data Mining defined as a process to find useful patterns and trends from large data sets [16]. Data mining done by finding hidden and valuable information from a chunk of data, producing a new information that could be useful.

There are 2 tasks that can be done by data mining, namely Predictive tasks and Descriptive tasks, that can be accomplished through data mining. Predictive tasks are intended to predict the value of a single attribute dependent on the values of other attributes [17]. Although the key objective of the descriptive task is to draw patterns that summarize the underlying associations in data (correlations, trends, groups, trajectories, and anomalies) [17].

The data that will be processed by data mining is not all structured and suitable for processing automatically by data miners. Because of the Big Data has a very large variety. Data analysts use a processing pipeline to solve this issue, where the raw data is gathered, cleaned, and transformed into a structured format. The data can be stored in a normal database system and subsequently analyzed using analytical methods to provide insights. [15].

2.4 Clustering

Clustering is an unsupervised model in which divides objects into 'k' natural groups called clusters. Clustering divides data collections into segments (natural grouping) whose members have similar characteristics [18]. According to Vijay, Clustering is the process of finding meaningful groups on data [19]. Unlike classification techniques that predict target data into a specific class, clustering is an approach that divides data into segments or groups that holds similarities. This approach is often used to see the differences clearly between groups or clusters [23].

Based on the above definitions, in brief clustering is a technique of partitioning data into several groups based on their characteristics. More precisely, the main task of clustering is to do data grouping in such way that the data points in the cluster have a common similarities between the data points in the same cluster [20]. For its implementation, the k-means algorithm is the most popular and simple approach, therefore in this study will be used k-means algorithm.

3. METHODOLOGY

This article study steps can be seen in Figure 1: Research of Methodology

3.1 Define Problem and Develop Objectives

On the first phase, problem identification conducted as an objective definition of the study done in this article, the problem identification include the process collecting the relevant data to the underlying problem issue, in this context which is COVID-19. Then, based on the

problem, the objective will be defined as an purpose of this study.

3.2 Literature Review

The second phase is the phase for finding references, in the form of research journals, papers, reference books, and other references that are used to form the basis of this research theory. The study conducted deals with the big data analytics implementation in the health sector.

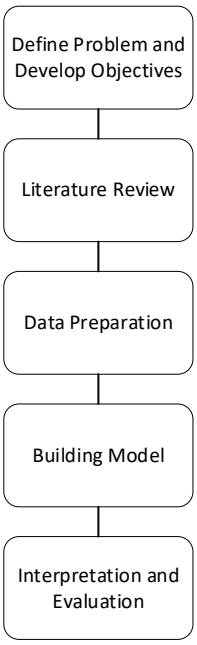

FIGURE 1: Research of Methodology [21]

3.3 Data Preparation

In the third phase, the preparation of the dataset that required during the study process is carried out. In this study, secondary data will be used as the basis for the study from third party platform called Kaggle [22]. Further explanation of the data will be discussed in the fourth section. In this process, only the necessary data were sorted during the clustering study process.

3.4 Building Model

In this phase, a model will be built that fits the purpose of the study which is clustering used to group the data that has been obtained into several segments or groups.

3.5 Interpretation and Evaluation

In the last phase, interpretation and evaluation of the model that has been made will be carried out. Interpretation aims to understand the context and results of the information generated by the model, and evaluation to test and develop existing models.

4. ANALYSIS AND DISCUSSION

4.1 Problem Understanding and Objectives Development

Before starting the data analysis phase, it is necessary to have an understanding of what problem to solve and form the purpose of this study. As previously presented, this study will be conducted an in-depth exploration of data on COVID-19 cases that have been collected over the past few months. One of the main problems with this pandemic condition is the increasing number of cases of COVID-19 patients every day. Based on the existing problems, it is proposed through this study to analyze a collection of case data from all over Indonesia to understand the existing conditions and provide a solution that can be implemented with big data analytics technology.

4.2 Data Analysis

Based on the data obtained, there are several columns representing each data type along with a name that already represents the information presented in each column. As attached to the description in previous section, we inspect each existing column; the total of the data used is 35 columns that each row representing the information regarding the cases. The data on COVID-19, which is presented as a reference, has a time-series format where there are many data points that represent cases on each date since the first case in Indonesia. The available scope of the data is between all over Indonesia, and data on cases based on each province in Indonesia. In this study, we will sort out and only use a few columns that are relevant to the purpose of this study.

J Popul Ther Clin Pharmacol Vol 30(4):e290–e300; 05 March 2023.

#	Column	Non-Null Count	Dtype
0	Location ISO Code	7797 non-null	object
1	New Cases	7797 non-null	int64
2	New Deaths	7797 non-null	int64
3	New Recovered	7797 non-null	int64
4	New Active Cases	7797 non-null	int64
5	Total Cases	7797 non-null	int64
6	Total Deaths	7797 non-null	int64
7	Total Recovered	7797 non-null	int64
8	Total Active Cases	7797 non-null	int64
9	Location Level	7797 non-null	object
10	City or Regency	0 non-null	float64
11	Province	7553 non-null	object
12	Country	7797 non-null	object
13	Continent	7797 non-null	object
14	Island	7553 non-null	object
15	Time Zone	7553 non-null	object
16	Special Status	1138 non-null	object
17	Total Regencies	7797 non-null	int64
18	Total Cities	7579 non-null	float64
19	Total Districts	7797 non-null	int64
20	Total Urban Villages	7577 non-null	float64
21	Total Rural Villages	7552 non-null	float64
22	Area (km2)	7797 non-null	int64
23	Population	7797 non-null	int64
24	Population Density	7797 non-null	float64
25	Longitude	7797 non-null	float64
26	Latitude	7797 non-null	float64
27	New Cases per Million	7797 non-null	float64
28	Total Cases per Million	7797 non-null	float64
29	New Deaths per Million	7797 non-null	float64
30	Total Deaths per Million	7797 non-null	float64
31	Case Fatality Rate	7797 non-null	object
32	Case Recovered Rate	7797 non-null	object
33	Growth Factor of New Cases	6990 non-null	float64
34	Growth Factor of New Deaths	6840 non-null	float64

In this study, we take advantage of a tool that has recently been popular and is often used in the data processing and analysis sector, namely python. As a programming language, Python provides a syntax that is fairly simple and easy to understand so that it can be used by ordinary people even in a short time. As a platform to make it easier to process, analyze data and display existing analysis results we use Jupyter Notebook to interact with the Python programming language during the process of analyzing data. To process and analyze data, we make use of several libraries provided by the community which are open-source. To perform data processing we use Pandas, to create a model we use Scikit-learn to do clustering, and finally we use Plotly to make a visualization of the model results so that it is easier to understand.

According to the study objectives, some information is needed to describe the conditions in a region, in the context of the COVID-19 pandemic, data that can represent how dangerous an area is the number of active cases at a time. Because based on an active case, it can represent how likely it is that the spread will increase if not handled. In addition, there are two factors that can affect the total active cases, namely when the patient is out of reach of active cases, there are only two possibilities, namely recovery or death. After that, to represent the number of accumulated cases that have occurred in an area, we decided to use the number of cases data for visualization. Through data analysis, the fields used in this study are date, location, total cases, total deaths, total recovered, total active cases.

TABLE 2: Data COVID-19 Indonesia from Kaggle

	Date	Location	Total Cases	Total Deaths	Total Recovered	Total Active Cases
1	3/1/2020	DKI Jakarta	489	20	39	430
2	3/2/2020	DKI Jakarta	491	20	39	432
3	3/2/2020	Indonesia	2	0	0	2
4	3/2/2020	Jawa Barat	12	5	4	3
5	3/3/2020	DKI Jakarta	493	20	39	434
6	3/3/2020	Indonesia	2	0	0	2
7	3/3/2020	Jawa Barat	13	6	4	3
8	3/4/2020	DKI Jakarta	495	20	39	436
9	3/4/2020	Indonesia	2	0	0	2
10	3/4/2020	Jawa Barat	14	6	4	4

Based on the fields that have been selected through the analysis process, here is a brief explanation of each field:

Date

Date when the number of COVID-19 cases was reported in a region Location or area where the COVID-19 case occurred

Total Cases

The accumulated number of COVID-19 cases in a region

Total Deaths

The number of deaths that have occurred due to COVID-19 in a region

Total Recovered

The number of patients who have recovered from COVID-19 infection

Total Active Cases

The number of cases currently active in a region

To prove the existence of a relationship between existing fields, correlation testing between variables was carried out using the correlation matrix.

	Total Cases	Total Deaths	Total Recovered	Total Active Cases
Total Cases	1.000000	0.977679	0.994801	0.941649
Total Deaths	0.977679	1.000000	0.962589	0.947538
Total Recovered	0.994801	0.962589	1.000000	0.902634
Total Active Cases	0.941649	0.947538	0.902634	1.000000

FIGURE 2: Correlation Matrix using Pandas

To provide an overview of the correlation between the selected fields, we use the Pandas library to show the correlation matrix between each variable. Figure 2: Correlation Matrix using Pandas shows that there is a correlation between total active cases, total deaths, total recovered, this happens because between every patient who recovers or dies will reduce new cases that occur. Then these three variables are part of the total cases in an area, so that the four variables above are indirectly related to one another. In this study, these four variables will continue to be used to visualize and form a model that is able to provide temporary solutions to prevent outbreaks that can cause a significant increase in cases. The correlation between each of the variables above can be seen from the correlation value above the number 0.9, which means that the relationship between them is very strong.

In Figure 3, you can see a diagram depicting the pandemic conditions in Indonesia since the day the COVID-19 case was first announced. As a basis, visualization based on data that represents COVID-19 as a whole in Indonesia illustrates a number of cases that continue to increase continuously in a number that is not small. The number of cases in August 2020 was in the position of around 109 thousand cases, but in the following month, September 2020, cases increased to 177 thousand and in early October 2020, the total cases in Indonesia reached 291 thousand. The total cases contained in the data presented do not include various cases that experienced delays from the laboratory and do not include the possibility of COVID-19 patients who did not test. Furthermore, apart from cases of death and recovery, only active cases were fluctuating in other words, continued to change, while in cases of death and recovery, it would

continue to increase as the number of cases occurred.

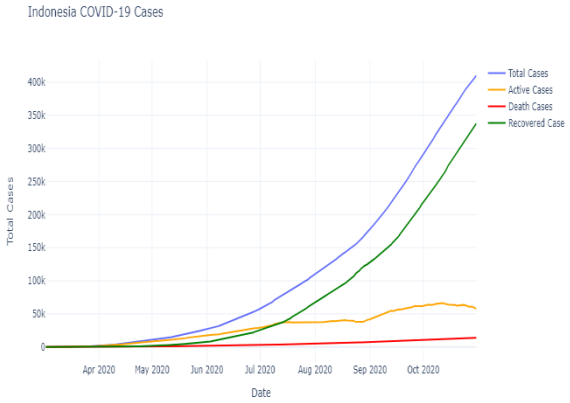

FIGURE 3: Indonesia COVID-19 Cases

In order to deal with a pandemic so that no future outbreaks will occur, a solution is needed that can directly monitor pandemic conditions in an area. This is required, especially in metropolitan cities that have high levels of citizen transmission. Of course, if direct monitoring is not carried out, COVID-19 cases will continue to increase and there will be more victims. As long as a vaccine solution has not been found, the only approach that can be taken is to implement social distancing and continuously monitor and track the progress of this pandemic case. Therefore, to support monitoring and tracking of coronavirus cases, a solution is needed that can provide information about the situation in an area.

One solution that can provide useful information for handling an epidemic condition like this is to group an area into groups based on the level of vulnerability of an area to the spread of COVID-19 at a certain point in time. In figure 4, a number of cases are compiled in each province where there are coronavirus cases. If sorted based on the number of cases, it can be seen that the five provinces with the largest cases are provinces with metropolitan capitals that have a high level of density and residents who keep moving from one place to another. Jakarta, which is the starting point for the spread of coronavirus and at the same time as a metropolitan city which is the capital of Indonesia, has the highest number of cases among other provinces.

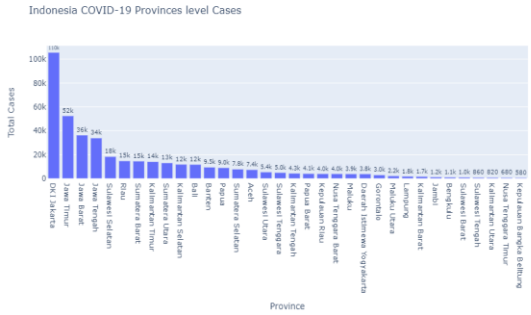

FIGURE 4: Indonesia COVID-19 Province Level Cases

4.3 Finding the most fit model

To form a label that can represent conditions in an area, an approach is needed that is able to determine the number of classes to be used. The most basic step for various types of unsupervised algorithms is to determine the optimal number of classes based on characteristics and experiments. One of the most popular methods for determining the number of clusters or groups is the Elbow Method. The Elbow Method approach is done by doing a number of iterations that try various combinations of the number of clusters. In each iteration, an inertia that is close to the intermediate position is taken. Inertia is an assumption of the distance of the data points on the cluster to the centroid. The purpose of doing the elbow method itself is to minimize the amount of distance between the data points and each centroid in the cluster. So, to find the optimal value, this study will perform iteration of clusters from 1 to 10 clusters.

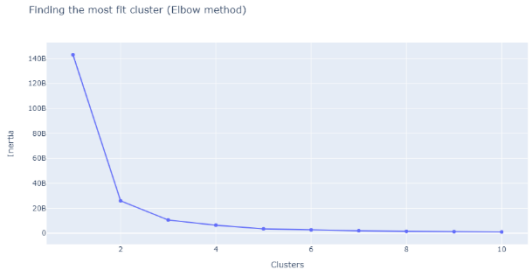

FIGURE 5: Model Fitting

To select the optimal number of clusters, a cluster with no significant decrease in the number will be selected in the next cluster. We process the cluster iteration until 10 clusters using Scikit-Learn library and give a clear picture, we use the Plotly library to create the visualization the

representation in Figure 5. In Figure 5, it can be seen that after the number of clusters exceeds 5, there is no significant difference, therefore the number of clusters that have a large decrease in inertia value will be taken compared to the previous cluster. From the Figure 5 above, it can be seen that after the number of clusters 2 to 3 has decreased quite significantly, but after that the decline was not that large. Therefore it can be concluded based on the figure above, the optimal number of clusters for data is 3 clusters.

4.4 Clustering

After determining the number of groups, the next process is to immediately apply the appropriate and optimal number of groups. Based on previously processed data, in this phase the data is adjusted to several fields that can represent the level of potential danger in an area or region. The selected fields are, Total Active Cases, Total Deaths, Total Recovered. These three fields were chosen because the number of cases that are still active in an area can give the possibility that further distribution can occur in an area. In addition, the number of deaths can provide an overview of the fatality rate in an area which may be influenced by external factors such as air and various other factors. And finally, Total Recovered, because until now the study on the behavior of the coronavirus pattern is still in the process of being researched, it is still not certain whether patients who have recovered are completely absent from the coronavirus and form immunity to the virus.

Then as a sample of provinces that might be the main model in representing the patterns and behavior of the coronavirus, we took DKI Jakarta as the model used in this clustering. Prior to clustering, the data was divided into two subsets, where the first subset containing 80% of the data points was used to train the clustering model, while the second subset which contained 20% of the remaining data was used to test the accuracy of the centroid points of the k-means clustering model used.

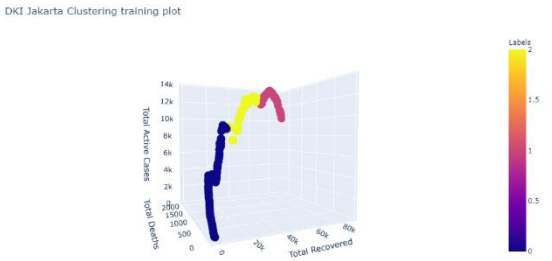

FIGURE 6: Clustering Training Plot

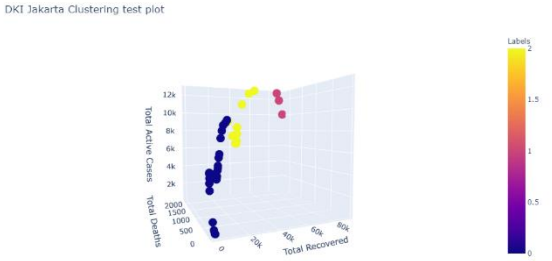

FIGURE 7: Clustering Test Plot

In Figure 6, clustering is done by utilizing training data to train the model in order to find suitable patterns as a basis for dividing groups based on existing data. Through figure 7, it is proven that by making a baseline model it can be used to classify a data point into a cluster that represents the level of risk of exposure to coronavirus in an area. After clustering and visualizing in the form of a 3-dimensional scatter plot, a category will be determined for the designation of each cluster based on the characteristics of each group:

High Risk

Moderate Risk

Low Risk

Based on this grouping, areas classified as high or medium risk areas are given a strict regulation and focused incentivized from the side of medical treatment in these areas. Whereas in low-risk areas it is handled by providing regulations that allow residents to go out if they comply with existing regulations such as wearing masks, maintaining cleanliness and carrying out social distancing, but it is still advisable to stay at home to prevent this possibility. Through the cluster visualization of the risk level above, it can be concluded that a representation that gives an idea

of how an area can be said to have a high risk, namely by having a large number of active cases, supported by the number of patients who died representing the fatality rate, and the number of patients recovered which describes how many patients who became victims who managed to live. It can be concluded, as in Table 3: Cluster Characteristic, that the higher the number of each variable in a region is able to represent the greater the risk of being exposed to the COVID-19 virus.

TABLE 3: Cluster Characteristic

	Total Active Cases	Total Recovered	Totah Deaths
0	Low	Low	Low
1	High	High	High
2	Moderate	Moderate	Moderate

4.5 Deployment

After the model has been obtained, this technology solution can be implemented by connecting or deploying the code that has been created through a cloud service that is connected to a telephone number database that can send situation information about the area where residents are located. This is intended to increase awareness about the conditions of the COVID-19 pandemic in a certain area. Furthermore, it is likely that this solution can be implemented into a smaller scope such as district or neighborhood. That way, residents can indirectly contribute to assisting the government in preventing an outbreak in an area in order to reduce the spread and improve existing conditions.

5. CONCLUSION

This study was carried out by utilizing a data mining approach technique on a large dataset of COVID-19 which continues to grow every day. In this study, the techniques used include correlation, clustering, and visualization to help understand the situation that occurs through large data. After analyzing and visualizing the number of coronavirus cases in Indonesian data, it can be concluded that the situation is still fluctuating and changing, especially as long as vaccines that pass the test are found. Therefore, this study analyzes the available data. Based on the analysis of existing data, we propose to take a clustering approach to group an area into a group that represents the conditions that occur in the area to assist the government and citizens in forming a regulation that is in accordance with existing conditions. Based on the number of cases, we took a sample of the city of DKI Jakarta as a training model. The result is a model with three clusters which can be classified into low risk, moderate risk, and high risk which represent the risk of exposure in an area.

REFERENCES

1. M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mob. Networks Appl., vol. 19, no. 2, pp. 171–209, 2014, doi: 10.1007/s11036-013-0489-0.
2. S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0217-0.
3. X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, “Data Mining with Big Data Xindong,” Ieeexplore.Ieee.Org, pp. 1–26, 2014, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6547630/.
4. O. Müller, I. Junglas, J. Vom Brocke, and S. Debortoli, “Utilizing big data analytics for information systems research: Challenges, promises and guidelines,” Eur. J. Inf. Syst., vol. 25, no. 4, pp. 289–302, 2016, doi: 10.1057/ejis.2016.2.
5. J. P. John and S. Vasudevan, “Big Data Analytics In Healthcare,” 2018 10th Int. Conf. Adv. Comput. ICoAC 2018, no. February, pp. 212–215, 2016.
6. W. Raghupathi and V. Raghupathi, “Big Data Analytics in Healthcare: Promise and Potential,” Heal. Inf. Sci. Syst., 2014, doi: 10.1186/2047-2501-2-3.
7. T. Singhal, “Review on COVID19 disease so far,” Indian J. Pediatr., vol. 87, no. April, pp. 281–286, 2020.
8. R. Madurai Elavarasan and R. Pugazhendhi, “Restructured society and environment: A review on potential technological strategies to control the COVID-19 pandemic,” Sci. Total Environ., vol. 725, p. 138858, 2020, doi: 10.1016/j.scitotenv.2020.138858.
9. T. J. Post, “Stay home, President says,” 2020.
10. S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, “Predicting

the growth and trend of COVID-19 pandemic using machine learning and cloud computing," Internet of Things, vol. 11, p. 100222, 2020, doi: 10.1016/j.iot.2020.100222.
11. K. Abdulmajeed, M. Adeleke, and L. Popoola, "Online Forecasting of Covid-19 Cases in Nigeria Using Limited Data," Data Br., vol. 30, p. 105683, 2020, doi: 10.1016/j.dib.2020.105683.
12. A. Mahmudan, "Clustering of District or City in Central Java Based COVID-19 Case Using K-Means Clustering," J. Mat. Stat. dan Komputasi, vol. 17, no. 1, pp. 1–13, 2020, doi: 10.20956/jmsk.v17i1.10727.
13. A. Mcafee and E. Brynjolfsson, "Spotlight on Big Data Big Data: The Management Revolution, 2012. Acedido em 15-03-2017," Harv. Bus. Rev., no. October, pp. 1–9, 2012.
14. V. Mayer-Schönberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, vol. 53, no. 9. Boston: Houghton Mifflin Harcourt, 2013.
15. C. C. Aggarwal and C. C. Aggarwal, Data Classification. 2015.
16. D. T. Larose and C. D. Larose, Data Mining and Predictive Analytics, 2nd ed. Wiley Publishing, 2015.
17. P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Pearson, 2013.
18. E. Turban, R. Sharda, and D. Delen, Business Intelligence and Analytics: Systems for Decision Support, Global Edition, 10th editi. Pearson Education Limited, 2014.
19. V. Kotu and B. Deshpande, Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner. Elsevier Science, 2014.
20. I. H. Witten, E. Frank, and M. a Hall, Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). 2011.
21. EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Wiley, 2016.
22. Hendratno, "COVID-19 Indonesia Dataset." 2020, doi: 10.34740/kaggle/dsv/1608770.
23. J. F. Andry, H. Tannady, G. D. Rembulan & A. Rianto. The importance of big data for healthcare and its usage in clinical statistics of cardiovascular disease. Journal of Population Therapeutics and Clinical Pharmacology, 29(04), 107-115, 2022.
24. D. Y. Heryadi, H. Tannady, G. D. Rembulan, B. Rofatin, & R. S. Sundari. Changes in behavior and welfare of organic rice farmers during the COVID-19 pandemic. Caspian Journal of Environmental Sciences, 21(1), 191-197, 2023.
25. J. F. Andry, L. Liliana, H. Tannady, & A. S. Arief. (2022, December). Data Centre Risk Analysis Using ISO 31000: 2009 Framework. In Journal of Physics: Conference Series (Vol. 2394, No. 1, p. 012032). IOP Publishing.
26. H. Tannady, J. M. Renwarin, A. N. D. Cora, & E. Purwanto. (2021, July). Production Planning and Inventory Control of Atonic Fertilizer Products Using Static Lot Sizing Method. In IOP Conference Series: Earth and Environmental Science (Vol. 819, No. 1, p. 012087). IOP Publishing.