Formosa Journal of Applied Sciences (FJAS)
Vol. 2, No. 7, 2023: 1689-1700

Analysis Clustering of the Global Pandemic Covid-19 using K-Means Algorithm

Bakti Siregar[1]*, Yosia[2]
Jurusan Statistika Universitas Matana
Corresponding Author: Bakti Siregar siregar.bakti@matanauniversity.ac.id

A R T I C L E I N F O

Keywords: Machine Learning, K-means Algorithm, Clustering Analysis, Global Covid-19

Received : 22, Mei
Revised : 24, June
Accepted: 26, July

A B S T R A C T

A pandemic such as Covid-19 is one of the biggest real problems ever in the world. This case has confirmed how uncertainty affects the global economy. The pandemic Covid-19 cannot be solved by one method over the world, it depends on the severity of the case. Therefore, this research aims to cluster the severity of Covid-19 using the K-means algorithm to reflect the global economic conditions using data sources from "Our World in Data". The results of this research can be used as materials to overcome the impact of the global pandemic by referring to policies and strategies from a country that is indicated in one cluster.

DOI prefik: https://doi.org/10.55927/fjas.v2i7.4877
ISSN-E: 2962-6447
https://journal.formosapublisher.org/index.php/fjas

INTRODUCTION

The Covid-19 pandemic was officially declared over by WHO on Thursday, May 4, 2023. However, the impact of this pandemic has a harming the economies of many countries in the world, especially in the part of tourism, small and medium enterprises, and investment (Sugihamretha, 2020). Several studies from various fields have declared some impact of the pandemic Covid-19 also induces sentiment of most investors toward the market tend to be negative (Lisbet, 2021). This condition is based on a sensitivity analysis that brings about a global economic decrease because of a major impact on international politics (Tanjung, S.I., 2021). In addition, the pandemic Covid-19 resulted in decreasing tax revenue and also increase some spending, which affected fiscal pressure, especially in lower-middle-income countries that required improvements to the tax system to reduce higher fiscal constraints (Hidayah et al, 2022). The resolution of this kind of economic effect can be actually resolved by learning about the experience how the Ebola virus in Liberia, due to increased public health spending, economic collapse, and decreased income due to the government's inability to increase revenue due to quarantine and curfews. (Nursyabany, I. 2022).

In general, efforts that can be made to tackle the spread of a type of pandemic like Covid-19 are by conducting quarantine, preparing public and personal health facilities, isolating infection cases, tracing and isolating all contacts exposed to the source of infection, involving public health resources, and providing financial support (Permadi & Sudirga, 2020). So that the impact of the pandemic Covid-19 in global economics also depends on the policy actions was taken by a country, such as implementing a travel limiting policy is one of the constraining factors in the relationship of mutual need between countries, besides that, there is a shortage of food supplies and medical equipment because countries prefer to hoard for their own interests (Kusno, F. 2020). The impact of the global economy is also exacerbated by the lack of manpower, transportation disruptions, closure of workplaces, restrictions on trade and travel, as well as the closing of land and air borders (Heninda et al., 2020).

The essence of problems as mentioned above, aims to apply the K-means algorithm to generate clustering conditions of Covid-19 around the world using the unsupervised learning approach. The results of this research can be used as a basis for making decisions for a country to determine the imposition of restrictions on land and air relations for countries indicated in the highest distribution clusters. Of course, this policy will really help maintain a balance in the interests of the supply of food, drink, medical equipment, and labor in a country. More than that, a country can also follow the regulations and policies of countries indicated in one cluster that has been successful in handling the pandemic.

LITERATUR REVIEW

According to (Barchitta et al., 2021), cluster analysis is a model of unsupervised learning that is usually used to group data based on certain characteristics. There are two algorithms that are most often used in cluster modeling namely; K-means and fuzzy C-means. These two algorithms have been proven to be able to group data very well based on similarities and dissimilarities (Askari, 2021). However, according to (Annas et al., 2022) K-Means has advantages over fuzzy c-means, namely; K-Means has the ability to form more homogeneous groups within one group, on the other hand, it is also able to form clear heterogeneity between groups, besides that the iterations used by K-Means will also stop at local optimum conditions. Research related to the clustering of Covid-19 cases in the world has been carried out by (Adha et al, 2021) applying the DBSCAN and K-Means algorithms. This paper applied the Silhouette Index (SI) to measure the value of cluster validity. However, researcher also shows a comparison of determining the number of clusters with three algorithms, namely; Elbow, Silhouette Index, and Gap Statistics.

Elbow Algorithm

The elbow algorithm is the oldest method used to determine optimum clusters for a set of data being analyzed, starting with testing 2 clusters, then adding one cluster to the maximum number of clusters to estimate the potential number of clusters, and in the end can determine the optimal number of clusters. (Shi et al., 2021). The number of clusters used in the k-means algorithm must be optimal so that the addition of k will not make a significant contribution (Jauhari et a., 2022). The k number of clusters is added one by one and the Sum Square Error (RMSE) value is recorded, where:

$$SSE = \sum_{k-1}^{k} \sum_{x_i \in S_k} \left|\left|X_i - C_k\right|\right|^2 \quad (1)$$

X_i is the i-th variable data, C_k is the k-th cluster center point, S_k is the k-th cluster and SSE is the sum of the average Euclidean Distance from all points to the center point. (Marutho et al., 2018) When the k value drops usually and forms a right angle, an indication of the optimal k value has been found.

Silhouette Index

The Silhouette index (SI) shows objects that should be in clusters and able to distinguish objects that should be in between clusters. (Wang & Xu, 2019) The SI obtains the optimal number of clusters when the difference between the average distance within clusters and the minimum distance between clusters, called the optimal clustering effect. The SI value is obtained from formula (2).

$$SI(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & if\ a(i) < b(i) \\ 0 & if\ a(i) = b(i) \\ \frac{a(i)}{b(i)} - 1 & if\ a(i) > b(i) \end{cases} \quad (2)$$

The SI values obtained can be interpreted as follows: 0.71 to 1.00 means that the data was perfectly divided very well; 0.51 to 0.70 means the data was accordingly divided well; 0.26 to 0.50 means the data was not appropriately divided; 0.00 to 0.25 means that the data was not clearly divided; (-0.01) to (-1.00) means that the data was incorrectly divided.

Gap Statistics Algorithm

The Gap Statistical Algorithm is the first algorithm proposed by (Tibshirani et al., 2001) to determine the optimal number of clusters in a data set with an unknown number of classifications. This algorithm genuinely applied sampling measurements like the Monte Carlo method to calculate the sum of the squares of the Euclidean distance between two measurements in each class contribution (Jauhari et a., 2022). Then, compared the grouping results to determine the optimal number of clusters. Mathematically write as (3).

$$\begin{aligned} Gap_n(k) &= E_n^*(\log W_k) - \log W_k E_n^*(\log(W_k) \\ &= \left(\frac{1}{p}\right)\sum_{b=1}^{P}\log(W_{kb}^*) \approx \left(\frac{1}{p}\right)\sum_{b=1}^{P}\log(W_{kb}^*)\, s(k) \\ &= \sqrt{\frac{1+P}{P}}\, s(k) \end{aligned} \tag{3}$$

Explanation:

k : number of clusters evaluated

W_k : dispersion within clusters for k clusters.

W_{kb} : dispersion within clusters for k clusters in data set refer to b

P : number of samples

$s(k)$: difference between ($\log(W_k)$) and ($\log(W_{kb})$)

Where $E_n^*(\log(W_k))$ is expectation of $\log(W_k)$ randomly generated by Monte Carlo. The value of k corresponds to the maximum value of Gap_k, it is the optimum number of k; satisfies when the minimum k of $Gap_k \geq Gap_{k+1} - S_{k+1}$. The most important thing that is unique in this study when compared to previous researchers is that the attributes used are more diverse and also the application of the Multicollinearity Test to select attributes that really have an impact on the cluster formation process. According to (Daoud, 2018), multicollinearity arises when two or more independent variables are correlated with each other. The existence of this multicollinearity often creates big problems in determining the influence factor of a variable. If the Variance Inflation Factor (VIF) value is not more than 10 then it means the tolerance value is not less than 0.1, then the model has no multicollinearity (A. Wulandari, 2021).

METHODOLOGY

This research uses a quantitative approach involving data on the spread of Covid-19 and global economic conditions sourced from the "Our World in Data" website. The first step is to ensure the missing value on any of the variables used. The second step is eliminating variables with VIF values above 10 to avoid multicollinearity. The third stage is selecting the optimal number of clusters using three method, called; elbow method, silhouette index, and Gap Statistics. The fourth stage, building a clustering model using the K-Means algorithm by applying the number of clusters obtained in the third step. The fifth stage,

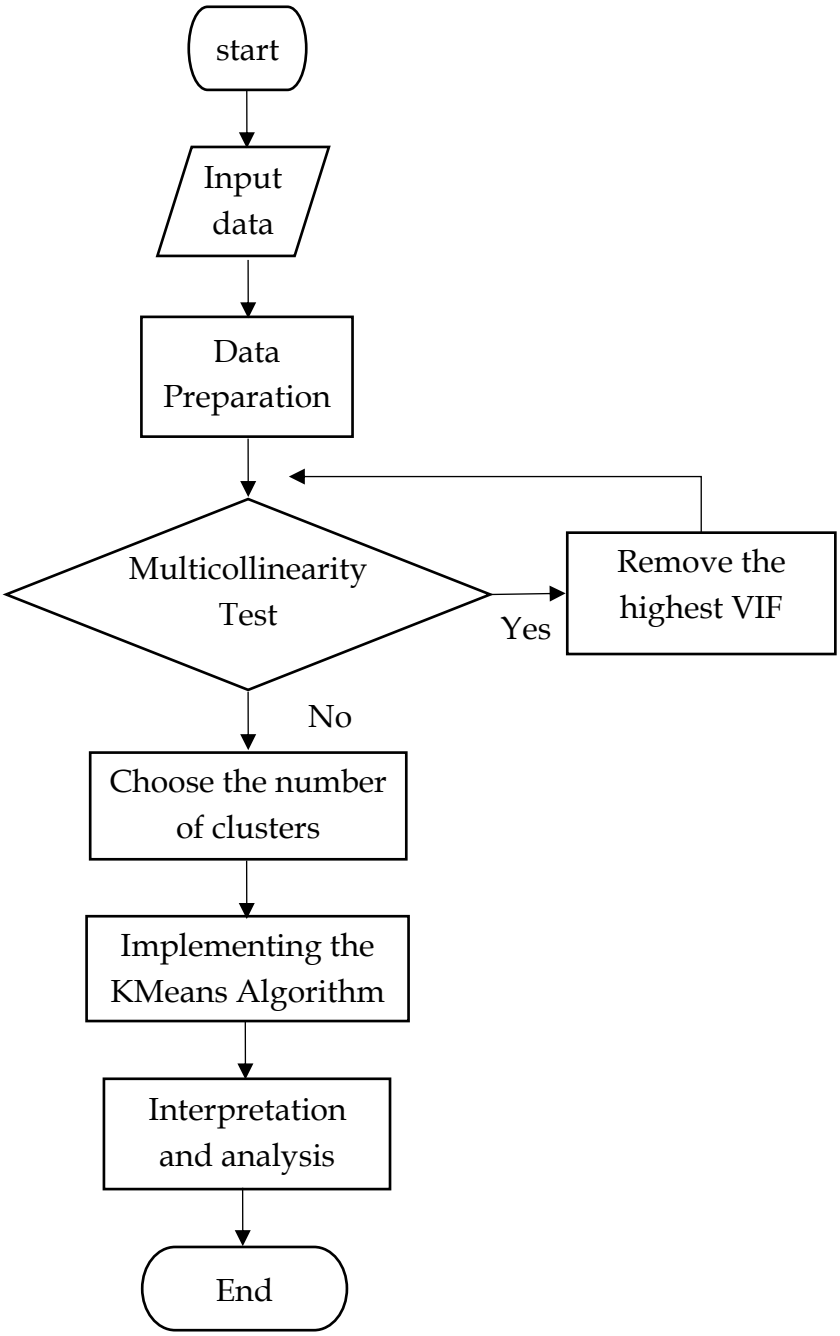

Figure 1. Research Flowchart

Visualizing the cluster to ensure that the data is well clustered. The sixth stage, analyzing global economic conditions based on the clusters found in the current condition of the spread of Covid-19. The final stage, concludes clustering analysis. In general, the methodology of this research are shown in Figure 1.

RESEARCH RESULT

The outcome of this research initiated from data preparation, multicollinearity test, search for the optimal number of clusters, application of the K-Means algorithm, and interpretation of the results. A complete discussion is shown as follows:

Data Preparation

The data used was obtained from the Our World in Data website, namely https://ourworldindata.org, in the form of Covid-19 case data for all countries in the world. The researchers used known data, namely the peak dates for the spread of COVID-19 cases from wave II (2022-01-26), wave III (2022-07-30), and wave IV (2022-12-27) globally refer to Figure 2. In addition, the researchers involved 29 numerical attributes which were considered cluster-forming variables as shown in Table 1. At the data preprocessing stage, a zero value was replaced for each missing data.

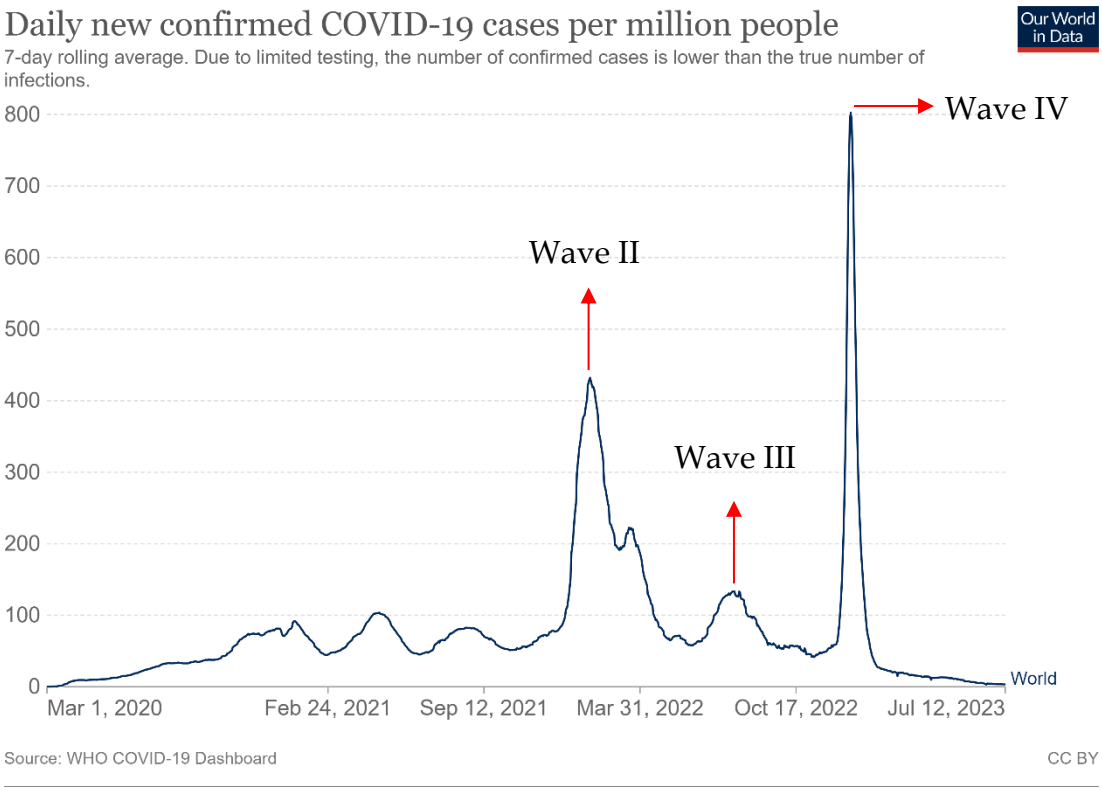

Figure 2. Daily confirmed Covid-19 (Wave II-IV)

Multicollinearity Test

This research does not have a dependent variable to be used as a reference for multicollinearity test calculations, so a dummy variable (y) is formed to help calculate the Variance Inflation Factor (VIF) value, see Table 1.

Table 1. VIF Score

	Variables	Tolerance	VIF
1	total_cases	0.0020670564	483.779740
2	new_cases	0.0034537800	289.537839
3	total_deaths	0.0050129119	199.484855
4	new_deaths	0.0050572460	197.736078
5	reproduction_rate	0.3667527100	2.726633
6	icu_patients	0.0494556132	20.220152
7	hosp_patients	0.0455037721	21.976200
8	total_tests	0.0561104063	17.822006
9	new_tests	0.0636531425	15.710143
10	positive_rate	0.6847316798	1.460426
11	tests_per_case	0.0094497661	105.822725
12	total_vaccinations	0.0003491720	2863.917789
13	people_vaccinated	0.0002912708	3433.230731
14	people_fully_vaccinated	0.0002530367	3951.995351
15	total_boosters	0.0033214550	301.072871
16	new_vaccinations	0.0016283086	614.134191
17	population_density	0.8176575808	1.223006
18	median_age	0.1482552139	6.745125
19	aged_70_older	0.1541926717	6.485393
20	gdp_per_capita	0.4531994824	2.206534
21	extreme_poverty	0.7640239894	1.308859
22	cardiovasc_death_rate	0.4607427822	2.170408
23	diabetes_prevalence	0.6952407899	1.438351
24	female_smokers	0.3809766930	2.624832
25	male_smokers	0.4285933999	2.333214
26	handwashing_facilities	0.6465200974	1.546742
27	life_expectancy	0.5708672958	1.751721
28	human_development_index	0.2135626753	4.682466
29	population	0.0132013921	75.749587

Based on the observations in Table 1, some variables have VIF value greater than 10. Therefore, the variable *total_cases* should be removed because it is the highest VIF value. Repeat this stage until there is no VIF value greater than 10.

Chose the Optimal Number of Cluster

Based on the observations in Figure 3, it is recognized the optimal number of k should be in 3 clusters. This optimal number of clusters has been found by using Euclidean distance algorithm, the formula is shown in equation (4) (Nishom, 2019).

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \quad (4)$$

Explanation:

d: the distance between x and y

x_i: attributes of data-i, $(i = 1, 2, 3, \dots, n)$

y_i: attribute centre of the -i cluster , $(i = 1, 2, 3, \dots, n)$

Implementation of K-Means Algorithm

The process of implementing clustering using the K-Means algorithm refers to the optimal number of clusters, namely 3 clusters. In this case, the researchers conducted a comparative analysis on clustering the spread of Covid 19 in waves I to wave IV globally. An example of implementing the K-Means algorithm clustering at the peak of the Covid-19 wave II outbreak is shown in Figure 4. The recapitulation of the cluster results in this study is shown in Table 2.

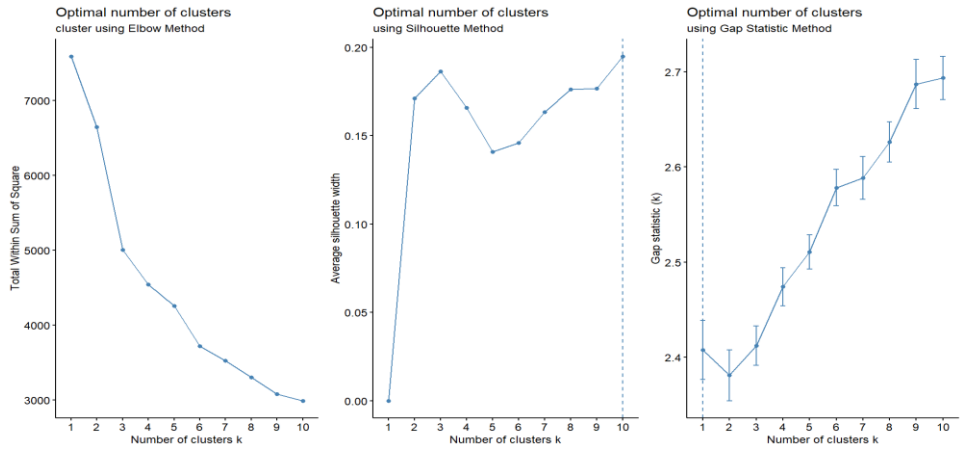

Figure 3. The Optimum Number of Cluster

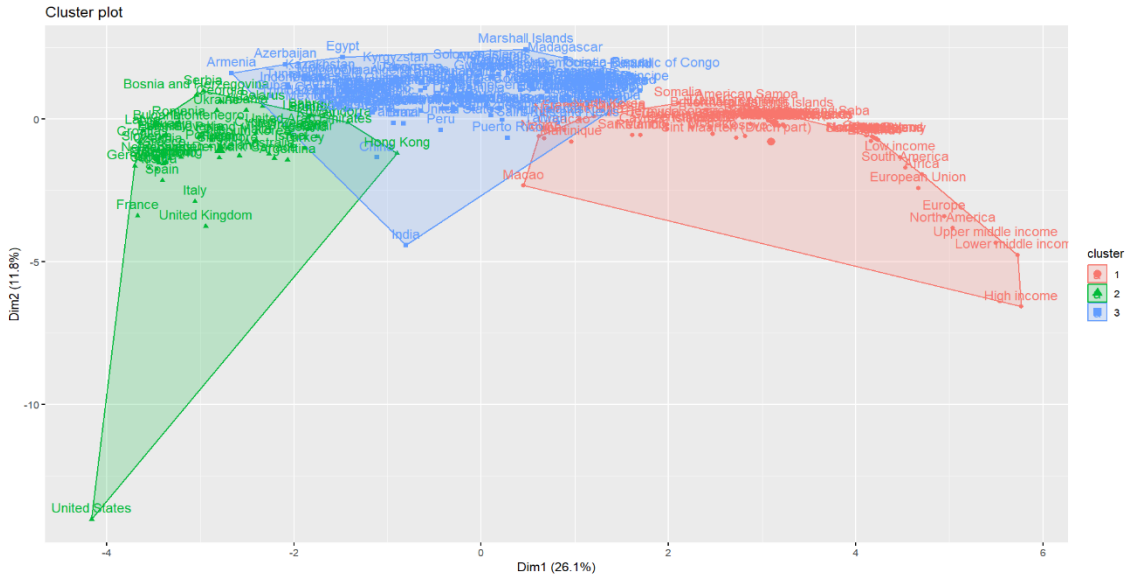

Figure 4. The clustering of Covid-19 in Wave II

Table 2. The Recapitulation of the Clustering

	wave2	wave3	wave4
Afghanistan	3	1	3
Africa	2	2	2
Albania	1	3	1
Algeria	3	1	3
American Samoa	2	2	2
Andorra	1	1	1
Angola	3	1	3
Anguilla	2	2	2
Antigua and Barbuda	3	1	3
Argentina	1	3	1
Armenia	3	1	3
Aruba	3	1	3
Asia	2	2	2
Australia	1	3	1
Austria	1	3	1
Azerbaijan	3	1	3
Bahamas	3	1	3
Bahrain	1	3	1
Bangladesh	3	1	3
Barbados	3	1	3
Belarus	1	3	1
Belgium	1	3	1
Belize	3	1	3
Benin	3	1	3
Bermuda	2	2	2
Bhutan	3	1	3
.....			
Vietnam	3	1	3
Wales	2	2	2
Wallis and Futuna	2	2	2
Yemen	3	1	3
Zambia	3	1	3
Zimbabwe	3	1	3

DISCUSSION

Based on data processing and analysis of research results that have been carried out, using the K-Means algorithm, the best cluster is obtained at k = 3. The clustering results have been validated using the Silhouette Index (SI). Based on the results of the cluster validation test on the results of clustering Covid-19 case data in the world using the K-Means algorithm, a multicollinearity test process is needed to ensure that the features used to form clusters are appropriate. In addition, it is necessary to search for a more optimal number of clusters by comparing several methods so that the selection can refer to the more dominant results. This is usually done when the optimal cluster recommendations found by one of the methods are inconsistent.

CONCLUSIONS AND RECOMMENDATIONS

The pattern from the research results can be used as a reference in describing the Covid-19 clustering model in the world. The cluster recommendations found can be used as strong reasons to determine whether a country should temporarily limit bilateral and economic relations with countries that are in an emergency cluster. Table 2 proves that there has been a change in clusters in each wave of the spread of Covid-19, which means that handling this case must be adjusted accordingly. More than that, if a country succeeds in

winning a problem, it can be used as a reference to be applied in dealing with the same problem.

ADVANCED RESEARCH

For further research, control tests for outlier data can be applied, adding cluster-forming features, as well as eliminating features that result in cluster bias by using more advanced methods.

ACKNOWLEDGMENT

Thank you to all members of Matana University, especially the Research Centre of Matana University who have been supporting my research in many aspects. In particular, to Matana University students, especially the statistics study program, who have been willing to help me carry out this Research.

REFERENCES

Adha, R., Nurhaliza, N., Sholeha, U., & Mustakim, M. (2021). Perbandingan algoritma DBSCAN dan k-means clustering untuk pengelompokan kasus Covid-19 di dunia. *SITEKIN: Jurnal Sains, Teknologi Dan Industri*, *18*(2), 206-211.

Annas, S., Poerwanto, B., & Sapriani, S. (2022). Implementation of K-Means Clustering on Poverty Indicators in Indonesia. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, *21*(2), 257-266.

Askari, S. (2021). Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. *Expert Systems with Applications*, *165*, 113856.

Barchitta, M., Maugeri, A., Favara, G., Riela, P. M., La Mastra, C., La Rosa, M. C., ... & Farruggia, P. (2021). Cluster analysis identifies patients at risk of catheter-associated urinary tract infections in intensive care units: findings from the SPIN-UTI Network. *Journal of Hospital Infection*, *107*, 57-63.

Hennida, C., Saptari, N. O., Aristyaningsih, I. G. A. A. R., & Febrianto, A. S. (2020). *Respons Negara Dan Institusi Global Terhadap Covid-19*. Airlangga University Press.

Hidayah, N., Yusuf, S. D., & Ajuna, L. H. (2022). STRATEGI KEBIJAKAN FISKAL DALAM MENGHADAPI DAMPAK PANDEMI COVID-19. *MUTAWAZIN (Jurnal Ekonomi Syariah)*, *3*(1), 28-39.

Jauhari, A., Anamisa, D. R., Mufarroha, F. A., & Suzanti, I. O. (2022, October). Grouping Madura Tourism Objects with Comparison of Clustering Methods. In *2022 IEEE 8th Information Technology International Seminar (ITIS)* (pp. 119-123). IEEE.

Kusno, F. (2020). Krisis Politik Ekonomi Global Dampak Pandemi Covid-19. *Anterior Jurnal*, *19*(2), 94–102. https://doi.org/10.33084/anterior.v19i2.1495

Lisbet. (2021). Penyebaran covid-19 dan Respons Internasional. Info Singkat Pusat Penelitian Dan Kajian DPR-RI.

Marutho, D., Hendra Handaka, S., Wijaya, E., & Muljono (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*, 533-538.

Nishom, M. (2019) "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," Jurnal Informatika: Jurnal Pengembangan IT, 4(1), pp. 20–24.Available at: https://doi.org/10.30591/jpit.v4i1.1253.

NURSYABANY, I. (2022). Peran United Nations International Children's Emergency Fund (Unicef) Terhadap Perlindungan Anak Akibat Wabah Virus Ebola Di Liberia Tahun 2014-2016.

Permadi, P. L., & Sudirga, I. M. (2020). Problematika Penerapan Sistem Karantina Wilayah Dan PSBB Dalam Penanggulangan Covid-19. *Jurnal Kertha Semaya*, *8*(9), 1355-1365.

Shang Y, Li H and Zhang R (2021) Effects of Pandemic Outbreak on Economies: Evidence From Business History Context. Front. Public Health 9:632043. doi: 10.3389/fpubh.2021.632043

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, *2021*(1), 1–16.

Tanjung, S. I. (2021). Dampak Covid – 19 Dalam Stabilitas Ekonomi Politik Internasional. *Ganaya : Jurnal Ilmu Sosial Dan Humaniora*, *4*(2), 654–671. https://doi.org/10.37329/ganaya.v4i2.1387

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering*, *569*(5). https://doi.org/10.1088/1757-899X/569/5/052024

Zakariah, M. A., Afriani, V., & Zakariah, K. M. (2020). *METODOLOGI PENELITIAN KUALITATIF, KUANTITATIF, ACTION RESEARCH, RESEARCH AND DEVELOPMENT (R n D)*. Yayasan Pondok Pesantren Al Mawaddah Warrahmah Kolaka.