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

Bakti Siregar1*, Yosia2
Jurusan Statistika Universitas Matana
Corresponding Author: Bakti Siregar siregar.bakti@matanauniversity.ac.id

ARTICLE INFO

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.
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.
LITERATURE 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 al., 2022). The \( k \) number of clusters is added one by one and the Sum Square Error (RMSE) value is recorded, where:

\[
SSE = \sum_{k=1}^{k} \sum_{x_i \in S_k} |X_i - C_k|^2
\]

\( X_i \) 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).

\[
SI(i) = \begin{cases} 
1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\
0 & \text{if } a(i) = b(i) \\
\frac{a(i)}{b(i)} - 1 & \text{if } a(i) > b(i) 
\end{cases}
\]
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 al., 2022). Then, compared the grouping results to determine the optimal number of clusters. Mathematically write as (3).

\[
Gap_n(k) = E_n^*(\log W_k) - \log W_k E_n^*(\log(W_k)) \\
= \left(\frac{1}{p}\right) \sum_{b=1}^{p} \log(W_{kb}^*) \approx \left(\frac{1}{p}\right) \sum_{b=1}^{p} \log(W_{kb}^*) s(k) \\
= \sqrt{\frac{1+p}{p}} s(k)
\]  

(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,

![Flowchart](image-url)

**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>
<thead>
<tr>
<th>Variables</th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>total_cases</td>
<td>0.0020670564</td>
<td>483.779740</td>
</tr>
<tr>
<td>new_cases</td>
<td>0.0034537800</td>
<td>289.537839</td>
</tr>
<tr>
<td>total_deaths</td>
<td>0.0050129119</td>
<td>199.484855</td>
</tr>
<tr>
<td>new_deaths</td>
<td>0.0050572460</td>
<td>197.736078</td>
</tr>
<tr>
<td>reproduction_rate</td>
<td>0.3667527100</td>
<td>2.726633</td>
</tr>
<tr>
<td>icu_patients</td>
<td>0.0494556132</td>
<td>20.220152</td>
</tr>
<tr>
<td>hosp_patients</td>
<td>0.0455037721</td>
<td>21.976200</td>
</tr>
<tr>
<td>total_tests</td>
<td>0.0561104063</td>
<td>17.822006</td>
</tr>
<tr>
<td>new_tests</td>
<td>0.0636531425</td>
<td>15.710143</td>
</tr>
<tr>
<td>positive_rate</td>
<td>0.6847316798</td>
<td>1.460426</td>
</tr>
<tr>
<td>tests_per_case</td>
<td>0.0094497661</td>
<td>105.822725</td>
</tr>
<tr>
<td>total_vaccinations</td>
<td>0.0003491720</td>
<td>2863.917789</td>
</tr>
<tr>
<td>people_vaccinated</td>
<td>0.0002912708</td>
<td>3433.230731</td>
</tr>
<tr>
<td>people_fully_vaccinated</td>
<td>0.0002530367</td>
<td>3951.995351</td>
</tr>
<tr>
<td>total_boosters</td>
<td>0.0033214550</td>
<td>301.072871</td>
</tr>
<tr>
<td>new_vaccinations</td>
<td>0.0016283086</td>
<td>614.134191</td>
</tr>
<tr>
<td>population_density</td>
<td>0.8176575808</td>
<td>1.223006</td>
</tr>
<tr>
<td>median_age</td>
<td>0.1482552139</td>
<td>6.745125</td>
</tr>
<tr>
<td>aged_70_older</td>
<td>0.1541926717</td>
<td>6.485395</td>
</tr>
<tr>
<td>gdp_per_capita</td>
<td>0.4531994824</td>
<td>2.206534</td>
</tr>
<tr>
<td>extreme_poverty</td>
<td>0.7640239894</td>
<td>1.308859</td>
</tr>
<tr>
<td>cardiovac_death_rate</td>
<td>0.4607427822</td>
<td>2.170408</td>
</tr>
<tr>
<td>diabetes_prevalence</td>
<td>0.6952407899</td>
<td>1.438351</td>
</tr>
<tr>
<td>female_smokers</td>
<td>0.3809766930</td>
<td>2.624832</td>
</tr>
<tr>
<td>male_smokers</td>
<td>0.4285933999</td>
<td>2.332141</td>
</tr>
<tr>
<td>handwashing_facilities</td>
<td>0.6465200974</td>
<td>1.546742</td>
</tr>
<tr>
<td>life_expectancy</td>
<td>0.57086272958</td>
<td>1.751721</td>
</tr>
<tr>
<td>human_development_index</td>
<td>0.2135626753</td>
<td>4.682466</td>
</tr>
<tr>
<td>population</td>
<td>0.0132013921</td>
<td>75.749587</td>
</tr>
</tbody>
</table>

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).

\[ d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \]  

(4)

Explanation:

\( d \): the distance between \( x \) and \( y \)

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

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


