e-ISSN: 2614-8404
p-ISSN: 2776-3234

K-Means Cluster Analysis of Sex, Age, and Comorbidities in the Mortalities of Covid-19 Patients of Indonesian Navy Personnel

Bambang Suharjo[1*)], Muhammad Satria Yuda Utama[2]
[1]Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[2]Master of Data Science and Business Analytic, Faculty of Computing, Asia Pacific University of Technology and Innovation
Email: [1]bambang_suharjo@tnial.mil.id, [2]satria@gmail.com

Abstract – **Covid-19 disease is still ongoing. It is necessary to do intensive research related to age, sex and congenital diseases so that management can be better planned. The research was conducted using data from Indonesian Navy personnel and their families, retired Indonesian Navy and their families. This study used k-means clustering for data grouping of Indonesian Navy personnel based on age, sex and congenital disease characteristics. The results of the k-means cluster clustering show that the k = 2 cluster has not been able to provide an explanation of the relationship between age, sex and comorbidity with the risk of death due to Covid-19. However, in the cluster with k = 3, it turns out that deaths due to Covid-19 are related to old age, men, even though there is no congenital disease. Meanwhile, using the k = 4 cluster, it is increasingly clear that deaths due to Covid-19 are closely related to old age, both men and women, with comorbidities.**

Keywords – comorbidity, Mortality, Covid-19, K-Means Cluster

I. INTRODUCTION

History records that pandemics have occurred many times and claimed millions of lives. There have been many significant pandemics and have caused enormous negative impacts in various fields including health, economy and even national security in the world [1]. Coronavirus disease 2019 (corona virus disease / Covid-19) is a new name given by the World Health Organization (WHO) for patients with the 2019 corona virus infection. The disease caused by the corona 19 virus was first reported from the city of Wuhan, China by the end of 2019. This positive single-strain RNA virus occurs by infecting the human respiratory tract. This virus is sensitive to heat and can be effectively inactivated by disinfectants containing chlorine. The source of the Covid-19 virus is thought to have come from animals, especially bats, and other vectors such as bamboo mice, camels and weasels. Common symptoms due to exposure to the Covid-19 virus include fever, cough and difficulty breathing. Clinical syndromes that appear after exposure to the Covid-19 virus can be grouped into uncomplicated, mild pneumonia and severe pneumonia [2], [3]. The spread is rapid throughout the world and the threat of a new pandemic until early 2021 has yet to be contained [4].

In Indonesia, exposure to the covid-19 virus was discovered in early March 2020. The development of sufferers due to this virus continues to increase until the beginning of 2021, reaching more than 1 million sufferers. Even though vaccines are starting to be found and used against this virus, various research is still needed to address the threat of this new virus pandemic as a whole.

Various studies that have been conducted in Indonesia and other countries show a link between age, sex and comorbidity to the dangers of the covid-19 virus on safety and recovery [2], [5]. Comorbidity is a patient congenital disease before exposure to a disease. In exposure to the disease caused by the Covid-19 virus, based on various studies, many comorbidities are believed to be dangerous for patient safety [6], [7], [8], [9], [10].

Comorbidity puts Covid-19 patients into vicious cycle of life and is strongly associated with significant morbidity and mortality. Comorbid individuals must adopt vigilant precautions and require careful management [6]. In the study, patients with the characteristics of old age, male, and critical illness were at increased risk of death compared to patients with other conditions at younger age, women and had no comorbidities Covid-19 patients with diabetes, chronic lung disease, cardiovascular disease, hypertension, HIV and other comorbidities may develop life-threatening situations [6], [7], [8], [9], [10].

This study will reveal the grouping of patients due to exposure to the Covid-19 virus in the Indonesian Navy with the characteristics of age, sex and comorbidity. Clustering can be used to partition data into groups, or clusters. A cluster can be described as a group of data objects that are more similar to other objects in their cluster than to data objects in other clusters [11], [12], [13].

Some previous research shown researches about clustering to analyze covid-19, age, gender, and comorbidities, such as:

The research aims to assess the relationship between sex, age, and comorbidity and mortality in Covid-2019 patients using clustering. The conclusion obtained is that gender, age, and comorbidities are partially related to the risk of death from Covid -19 [14]. Next research on the use of the k-means clustering of covid data obtained from Kaggle resulted in clusters with different levels of sufferers and deaths. With these results, it is recommended to carry

e-ISSN: 2614-8404
p-ISSN: 2776-3234

out different treatments in areas in different clusters [15]. Another study conducted grouping of districts and cities in Central Java, Indonesia based on Covid -19 cases using k-means clustering. The results show that two of the 3 clusters are areas that must be considered by the government because they are areas with a high number of active cases and high cases of Covid -19 deaths [16].

In another study, Artificial Neural Networks and k-means cluster were used on data on the current situation of the spread of Covid-19 in Indonesia for clustering. The resulting clusters consist of many provincial clusters in Indonesia with groupings on the characteristics of positive growth, recovery and death [17], [18], [19].

Based on these previous studies, it appears that clustering of Covid-19 sufferers is important. Thus, it needs to be made more comprehensive by adding patient comorbidities to the analysis. The results of these research are expected to be able to reveal more detailed characteristics about sufferers of exposure to the covid-19 virus. This is important to do to support the current pandemic policy model and its future anticipation.

II. RESEARCH METHODOLOGY

A. Data

The data used in this study were Covid-19 patient data, including: Indonesian Navy personnel and their families, retired members of the Indonesian Navy and their families. Data collected from March 2020 to December 2020. Data was collected using the census method, covering all of these data on those who suffer from Covid-19 which were obtained from the Indonesian Navy Health Service database.

B. Research Steps

The research steps were carried out as follows:

1. Data is collected from the Indonesian Navy covid-19 database
2. Cleaning data by eliminating the variables of the patient's name, unit origin, and the relationship with members of the Indonesian Navy if the family.
3. Data transformation
4. Create a descriptive analysis
5. Choose the best number of clusters using the elbow method.
6. Perform clustering calculations using the k-means method for the best clusters according to the elbow method
7. Analysis of the characteristics of the resulting clusters
8. Perform a comparison analysis of the characteristics of the clustering analysis results.

C. Flowchart Diagram

Based on the research steps above, the research flowchart was compiled as follows.

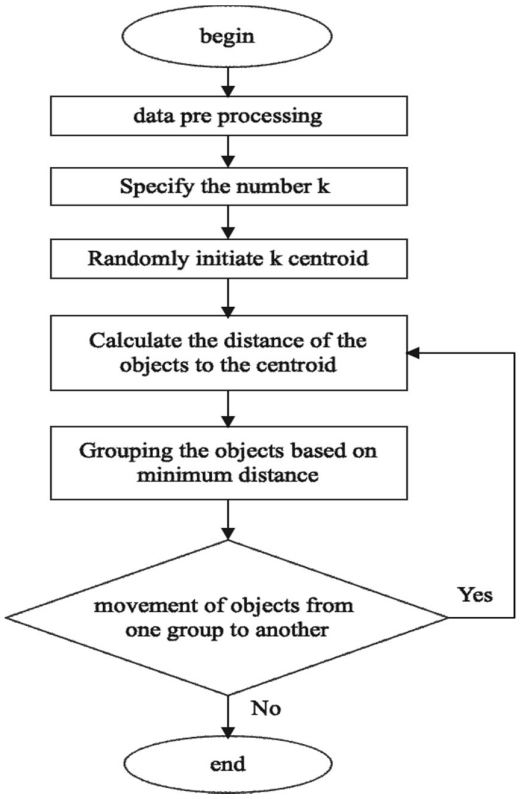

Figure 1. Flowchart diagram of the research

III. RESULTS AND DISCUSSION

A. Descriptive of Indonesian Navy Covid-19 Patients

Data on patients with Covid-19 as a whole can be described starting from gender (male and female), results of hospital treatment or independent isolation (recovering and dying), comorbidity (heart, lung and diabetes), presented in Figure 2, as following.

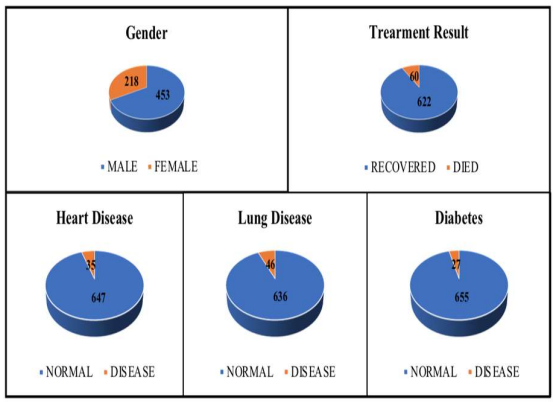

Figure 2. Distribution of Indonesian Navy Covid-19 sufferers based on gender (male and female), treatment results, and comorbidities.

Based on the pie chart above, it appears that the number of Indonesian Navy sufferers of Covid-19 from

e-ISSN: 2614-8404
p-ISSN: 2776-3234

March 2020 to December 2020 reached 682 people, with the breakdown of men = 453 people, 218 women. Meanwhile, 622 people were recovered from the treatment and 60 people died. The co-morbidities of the sufferers included: 35 heart disease, 46 lung disease and 27 diabetes. In addition, the age of the patients was between 4 years old until 88 years old and the mean was 36.8 years.

B. Cluster Analysis using k-mean

To get the size of the number of good clusters, it is necessary to select k. One way of selecting k can be done with the elbow method. The pseudocode for selecting k with Python language is as follows.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from kneed import KneeLocator
from sklearn.cluster import KMeans
sse = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(TRANSF_DATA)
    sse.append(kmeans.inertia_)
plt.plot(range(1, 11), sse, marker="o", color="red")
plt.title('Elbow METHOD')
plt.xlabel('Number of clusters')
plt.ylabel('SSE')
plt.show()
kl = KneeLocator(range(1, 11), sse, curve="convex", direction="decreasing")
print()
print('BEST K USING ELBOW METHOD IS','"', kl.elbow, '"')
```

Figure 3. Pseudocode of elbow method

The output results from the pseudocode above, then the elbow plot can be presented in accordance with the figure 4, as follow

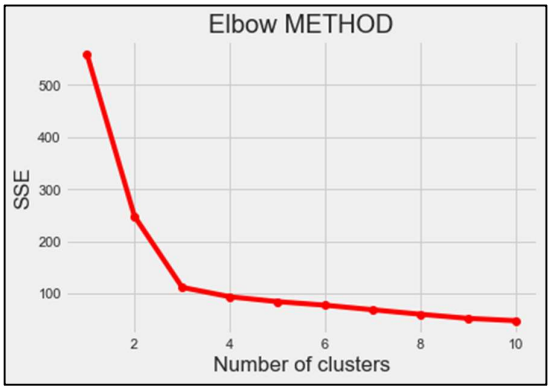

Figure 4. Elbow plot of best number of clusters

From the elbow plot, it can be seen that the elbows are at k = 2 to 4. Thus, the three options (k = 2, k = 3, and k = 4) are a priority for clustering calculations. Furthermore, the pseudocode used to carry out the clustering process using the k-means cluster method and plot the clustering results on the number of clusters 2, 3 and 4 which are carried out using Python, as follows

```
X = x_scaled
kmeans = KMeans(n_clusters = 2, init = 'k-means++', random_state = 42)
y_kmeans = kmeans.fit_predict(X)
print(y_kmeans)
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 300, c = 'black',
            label = 'Centroids')
plt.title('CLUSTERS (K=2) OF INDONESIAN NAVY COVID 19')
plt.xlabel('CRITERIA')
plt.ylabel('RESULT')
plt.legend()
plt.show()
```

Figure 5. Pseudocode of clustering plots

The output result of pseudocode in Figure 6 can be presented as follows.

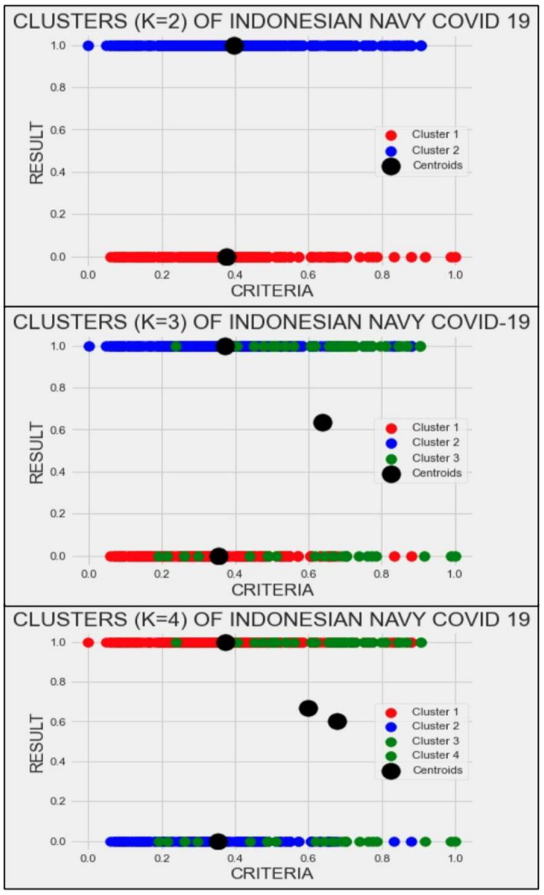

Figure 6. Plot of Objects clustering

Using centroid of every cluster, we can conclude that the characteristic of every kind of clustering, can be shown in table 1,2, and 3 below.

Table 1. Characteristic of Covid-19 of Indonesian navy personnel using 2 clusters

	Characteristic	**Percentage of member**
Cluster 1	Younger, dominated by female, dominated by normal heart, dominated by normal lung, and dominated by normal diabetes, and dominated by recovered.	32.7%
Cluster 2	Older, dominated by male, dominated by normal heart, dominated by normal lung, and dominated by normal diabetes, and dominated by recovered.	67.3%

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Based on the description of the characteristics of Indonesia n Navy personnel exposed to covid-19, young or old, most recovered and a small proportion died. However, it cannot be concluded that comorbid disorders are associated with mortality. So that clustering with cluster= 2 has not been a ble to explain in detail the comorbidity and mortality due t o covid-19.

Table 2. Characteristic of Covid-19 of Indonesian navy personnel using 3 clusters

	Characteristic	**Percentage of member**
Cluster 1	Young, dominated by women, dominated by normal heart, dominated by normal lungs, normal and dominated by normal diabetes, and recovered	29.4%
Cluster 2	Older, male, dominated by normal heart, dominated by normal lungs, normal and dominated by normal diabetes, and recovered	61.6%
Cluster 3	Old, dominated by men, dominated by normal heart, dominated by normal lungs, donimated by normal diabetes and died	8.8%

The results of the clustering according to table 2 show that a high risk of death is closely related to old age and men, with or without comorbidities. The conclusion from clustering into 3 clusters has not shown a detailed conclusion. Thus, it needs to be developed into 4 clusters.

Table 3. Characteristic of Covid-19 of Indonesian navy personnel using 4 clusters

	Characteristic	**Percentage of member**
Cluster 1	Young, male, dominated by no congenital disease, recovered	61.7%
Cluster 2	Young, female, dominated by no congenital disease, recovered	29.5%
Cluster 3	Elderly, male-dominated, some have cardiac comorbidities, mostly pulmonary comorbidities, no diabetic comorbidities, died	4.4%
Cluster 4	Elderly, predominantly male, some have cardiac comorbidities, no pulmonary comorbids, some have diabetes comorbidities, died	4.4%

Based on the clustering in table 3 which contains 4 clusters, it appears that the risk of death from covid-19 will be high if the elderly, male or female, with cardiovascular, lung and diabetes comorbidities. Therefore, it is necessary to be more careful, more intensive treatment in patients with these characteristics.

IV. CONCLUSION

Based on the results of the k-means cluster clustering, it appears that the k = 2 cluster has not been able to provide an explanation of the relationship between age, sex and comorbidity with the risk of death due to covid-19. However, in clusters with k = 3, it appears that deaths from covid-19 are related to older age, men, even though there is no congenital disease. Meanwhile, using the k = 4 cluster, it is increasingly clear that deaths from covid-19 are closely related to older age, both men and women, with comorbidities.

REFERENCES

[1] Qiu, W., Rutherford, S., Mao, A., & Chu, C. (2017). The Pandemic and its Impacts. *Health, Culture and Society*, *9*, 1–11. https://doi.org/10.5195/hcs.2017.221

[2] Yuliana. (2020). Corona virus diseases (Covid -19); Sebuah tinjauan literatur. *Wellness and Healthy Magazine*, *2*(1), 187–192.

[3] Anjorin, A. A. (2020). The coronavirus disease 2019 (COVID-19) pandemic : A review and an update on cases in Africa The coronavirus disease 2019 (COVID-19) pandemic : A review and an update on cases in Africa. *Asian Pacific Journal of Tropical Medicine*, *13*(April). https://doi.org/10.4103/1995-7645.281612

[4] Handayani, D., Hadi, D. R., Isbaniah, F., Burhan, E., & Agustin, H. (2020). Penyakit Virus Corona 2019. *Jurnal Respirologi Indonesia*, *40*(2), 119–129.

[5] Putri, R. N. (2020). Indonesia dalam Menghadapi Pandemi Covid-19. *Jurnal Ilmiah Universitas Batanghari Jambi*, *20*(2), 705. https://doi.org/10.33087/jiubj.v20i2.1010

[6] Filardo, T. D., Khan, M. R., Krawczyk, N., Galitzer, H., Karmen-Tuohy, S., Coffee, M., Schaye, V. E., Eckhardt, B. J., & Cohen, G. M. (2020). Comorbidity and clinical factors associated with COVID-19 critical illness and mortality at a large public hospital in New York City in the early phase of the pandemic (March-April 2020). *PLoS ONE*, *15*(11 November), 1–16. https://doi.org/10.1371/journal.pone.0242760

[7] Ejaz, H., Alsrhani, A., Zafar, A., Javed, H., Junaid, K., Abdalla, A. E., Abosalif, K. O. A., Ahmed, Z., &

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Younas, S. (2020). COVID-19 and comorbidities: Deleterious impact on infected patients. *Journal of Infection and Public Health*, *13*(12), 1833–1839. https://doi.org/10.1016/j.jiph.2020.07.014

[8] Gold, M. S., Sehayek, D., Gabrielli, S., Zhang, X., McCusker, C., & Ben-Shoshan, M. (2020). COVID-19 and comorbidities: a systematic review and meta-analysis. *Postgraduate Medicine*, *132*(8), 1–7. https://doi.org/10.1080/00325481.2020.1786964

[9] Kun'ain, U. I. A., Rahardjo, S. S., & Tamtomo, D. G. (2020). Meta-Analysis: The Effect of Diabetes Mellitus Comorbidity on the Risk of Death in Covid-19 Patients. *Indonesian Journal of Medicine*, *5*(4), 368–377. https://doi.org/10.26911/theijmed.2020.05.04.12.

[10] Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., Ji, R., Wang, H., Wang, Y., & Zhou, Y. (2020). Prevalence of comorbidities and its effects in coronavirus disease 2019 patients: A systematic review and meta-analysis. *International Journal of Infectious Diseases*, *94*, 91–95. https://doi.org/10.1016/j.ijid.2020.03.017

[11] Frigui, H. (2008). Cluster Analysis: Basic Concepts and Algorithms. In *2008 1st International Workshops on Image Processing Theory, Tools and Applications, IPTA 2008*. https://doi.org/10.1109/IPTA.2008.4743793

[12] Kumar, M., & Verma, A. (2018). Clustering Techniques - A Review. *International Journal of Computer Sciences and Engineering*, *6*(6), 1091–1099. https://doi.org/10.26438/ijcse/v6i6.10911099

[13] Thrun, M. (2018). Approaches to Cluster Analysis. In *Projection-Based Clustering through Self-Organization and Swarm Intelligence* (pp. 21–31). https://doi.org/10.1007/978-3-658-20540-9_3

[14] Biswas, M., Rahaman, S., Biswas, T. K., Haque, Z., & Ibrahim, B. (2021). Association of Sex, Age, and Comorbidities with Mortality in COVID-19 Patients: A Systematic Review and Meta-Analysis. *Intervirology*, *64*(1), 36–47. https://doi.org/10.1159/000512592

[15] Indraputra, R. A., & Fitriana, R. (2020). K-Means Clustering Data COVID-19. *Jurnal Teknik Industri*, *10*(3), 275–282.

[16] Mahmudan, A. (2020). Clustering of District or City in Central Java Based COVID-19 Case Using K-Means Clustering. *Jurnal Matematika, Statistika Dan Komputasi*, *17*(1), 1–13. https://doi.org/10.20956/jmsk.v17i1.10727

[17] Khotimah, T., & Darsin. (2020). Clustering Perkembangan Kasus Covid-19 di Indonesia Menggunakan Self Organizing Map. *Jurnal Dialektika Informatika (Detika)*, *1*(1), 23–26.

[18] R, R. P., & E, Y. A. (2020). *Analisis Cluster Virus Corona (COVID-19) di Indonesia pada 2 Maret 2020 – 12 April 2020 dengan Metode K-Means Clustering*. *May*, 1–6.

[19] Virgantari, F., & Faridhan, Y. E. (2020). K-Means Clustering of COVID-19 Cases in Indonesia s Provinces. *Proceedings of the International Conference on Global Optimization and Its Applications*.