Clustering Indonesian Provinces by Disaster Intensity using K-Means Algorithm: a Data Mining Approach

Wibowo A.[1*], Rohman N.[1], Rusdah[1], Achadi A.H.[1] and Amri I.[2]
1. Universitas Budi Luhur, Jakarta, INDONESIA
2. Universitas Gadjah Mada, Yogyakarta, INDONESIA
*arief.wibowo@budiluhur.ac.id

Abstract
Indonesia, as part of the Pacific Ring of Fire, is highly susceptible to various natural hazard events including geological and hydrometeorological disasters. This study aims to classify Indonesian provinces based on their disaster vulnerability using the K-Means clustering algorithm. The clustering process considers area and population data to ensure accurate grouping.

The results of the study successfully categorized Indonesian provinces into distinct clusters based on their disaster vulnerability levels. These clusters provide valuable insights for the National Disaster Management Agency (BNPB) and local governments in prioritizing disaster preparedness and response efforts. By identifying regions with higher vulnerability, this research contributes to the development of more targeted and effective disaster mitigation strategies, ultimately helping to reduce potential loss of life and property.

Keywords: Disaster intensity, natural hazards, K-Means, clustering, data mining.

Introduction

Disasters are extraordinary events that can threaten and affect people's lives and livelihoods. They are caused by natural or anthropogenic factors that affect the environment and society. Disasters are caused by various activities that damage natural objects on the face of the earth. The impact of this incident caused various problems including disruption to human activities. In addition, most disasters damage essential objects in the vicinity such as houses, public facilities and workplaces. Disasters also pose a high risk of loss of life including injury and death.

The risk of disasters is also exacerbated by unexpected or sudden events[17]. This has a severe impact on humanity and even causes death like an earthquake caused by pressure due to moving plates. Flooding was caused by unauthorized felling of trees and heavy rain. Landslides are caused by converting forests into residential land, buildings etc.[21]

Indonesia, as part of the Pacific Ring of Fire, frequently experiences disasters such as earthquakes, volcanic eruptions and tsunamis which have significant economic and social impacts on the affected regions. These disasters have led to substantial macroeconomic losses including a notable reduction in regional GDP and disruptions to labor markets across various provinces. This underscores the critical need for enhanced disaster preparedness and mitigation strategies to minimize the economic fallout of such events[9].

Indonesia is a country that is prone to hydrometeorological disasters, namely disasters caused by climate and weather changes. According to Law Number 24 of 2007, a disaster is an event or a series of events caused by natural and non-natural factors or human factors that threaten and disrupt people's lives and livelihoods, resulting in losses for humans, environmental damage, property loss and psychological impact. Given the severe macroeconomic losses caused by disasters, as evidenced by the significant reduction in regional GDP across various provinces, it is crucial for Indonesia to enhance its disaster preparedness and mitigation strategies[9]. These disasters occur because of threats and vulnerabilities that society is unable to overcome, threatening all regions of Indonesia including land, mountains and coasts[5].

Data mining is extracting information from large data sets to identify hidden knowledge that can be used in real-time. The data mining process involves various data analysis algorithms including Clustering, Association and Classification, as well as other techniques[8]. Clustering is an activity (task) that aims to group similar and different data into clusters or groups. This results in the maximum possible data similarity within a cluster and minimum data similarity between clusters. Clustering can also be interpreted as a data segmentation technique applied in various fields such as marketing, business problem analysis, market segmentation and forecasting, computer vision patterns and regional zoning to object identification and image processing[1]. Clustering can also be used as a grouping to collect data with known class labels[16].

From the explanation of clustering, to find out whether Indonesia is vulnerable or very vulnerable to disasters, a method is needed that can group data about disasters in Indonesia. The method used by researchers in this research is the K-Means method.

One method that can be used to group disaster data is the K-Means method. The K-Means method was once used entitled "Application of the K-Means clustering algorithm to group provinces based on the number of villages/subdistricts with efforts to anticipate/mitigate disasters"[13]. This explains that a disaster is an event or damage caused by nature that generally occurs suddenly or unexpectedly and can cause

**** Author for Correspondence***

material loss and even loss of life. Another research is entitled "Application of the K-Means Clustering Algorithm to Determine Priority for Recipients of Housing Assistance Due to Disasters"[10].

Here, K-means identifies four priority levels for recipients of disaster housing assistance, namely Very Priority (C0), Priority (C1), Low Priority (C2) and No Priority (C3). Similar research was conducted with the title "Application of the K-Means Method in Clustering Earthquake Prone Areas in Indonesia," which produced a Silhouette index value of 0.3245, so the optimal number of clusters is $k = 6$[4]. The highest earthquake was 5,125 events in one cluster and the lowest was 632 earthquake events in another cluster with the number of clusters $k = 6$.

Our study, which applies the K-Means method to identify patterns of disaster events in Indonesia, has the potential to provide valuable insights for your work. By grouping disaster data, we can identify consistent patterns that can serve as the basis for further analysis and the design of more targeted and efficient mitigation strategies. This research, therefore, offers a deeper understanding of the nature and patterns of disasters in Indonesia, contributing to the development of disaster risk reduction and management.

Review of Literature

This research applies a literature review method to study the grouping of disaster events in Indonesia based on disaster type using K-means. The literature review phase includes four steps: formulating the problem, searching the literature, evaluating the data and analyzing and interpreting the information. The formulation process begins with selecting a research topic, focusing on data mining and big data analysis related to disasters in Indonesia. A literature search was conducted using tools such as Publish and Perish, using the keywords "data mining" and "K-means algorithm" with information from Google Scholar, Sprott and Scopus. Data evaluation involves categorizing journal articles according to their relevance and topicality. The ten selected articles were then reviewed using similarity search, comparison, criticism, synthesis and summary techniques. A literature search on the topic of data mining and big data analysis shows a variety of interesting findings from 10 papers related to this topic.

Based on a literature review, this research uses the K-means clustering algorithm to overcome various aspects of disasters in Indonesia. The focus is on analyzing and grouping events based on type, prioritizing post-disaster housing assistance, understanding event patterns by considering certain factors, estimating vulnerable locations and K-in grouping landslide-prone areas. The aim is to improve understanding, risk management and response to regional disasters.

This research also requires relevant theory about data mining and its algorithms. Data mining involves using pattern recognition technology, statistics and mathematics to identify new patterns, trends and relationships in large amounts of data. The aim is to identify essential patterns from an extensive database to support future decision-making[15]. Clustering is the process of grouping data into two or more groups based on the similarities between the data in these groups. This method is part of unsupervised learning and does not require labeling each group. Clustering is used to group data between groups of data based on similar attributes[19].

K-Means is a clustering algorithm that divides data into several groups based on the closest centroid. The goal is to group data to maximize similarities within groups and minimize differences between groups. This process includes determining the number of clusters, initializing the centroid, grouping the data and updating the centroid value until convergence[13].

Material and Methods

The data was gathered from the 2023 Statistics Indonesia (BPS) dataset and supplemented with additional sources. This study employs a data mining approach, specifically CRISP-DM (Cross-Industry Standard Process for Data Mining), which is a commonly used process model in research for solving problems[7]. The research process involves six phases of CRISP-DM as illustrated in figure 1.

Business understanding: During this phase, the main focus is to understand the goals and analyze business needs which are then translated into strategies. In this particular case, the research objective was to determine the severity of disasters in Indonesia and group the provinces based on these characteristics. The focus of the disasters is more related to natural hazard events. The outcome of this research will provide an overview of disaster risk reduction in Indonesia which can be used by the Government and related institutions for better planning and management. Additionally, this research will provide valuable insights for local communities and researchers.

Data understanding: Data understanding is carried out by comprehensively understanding the data requirements needed to solve the problems in this research[11]. This stage is the data collection stage or the data collection process. This research uses data from the Statistics Indonesia (BPS) and the data collected is about disasters in Indonesia in 2023.

Data preparation: At the data preparation stage, we used several steps contained in KDD to process the data taken. KDD is an abbreviation for Knowledge Discovery in Database. Several stages of KDD used in this research are data cleaning, data integration, data selection and data transformation[12].

Modeling: Data mining techniques are directly involved in the modeling stage of this research. The first step in this phase is to select and apply appropriate modeling techniques and adjust the model rules to optimize results.

Table 1
Research literature.

No	Author	Title	Key Remark(s)
1	Mariam et al[10] (2023)	Application of the K-Means Clustering Algorithm to Determine Priorities for Recipients of Home Assistance Due to Disasters	This research demonstrates the application of the K-means clustering algorithm to disaster aid recipients, using open data from West Java Province on house damage across 27 regencies/cities.
2	Wahidah et al[8] (2023)	Grouping Disaster Prone Areas in Jember Regency Using the K-Means Clustering Method	This research uses the K-Means method to analyze and group disaster-prone areas in Jember Regency based on the Indonesian disaster risk index, aiming to support local government in disaster management. The findings will aid in better understanding and addressing disaster risks in the region.
3	Amaliah et al[2] (2023)	Grouping Disaster Data by Region Using the K-Means Algorithm	This study identifies areas with the highest disaster frequency and diversity to provide deeper insights into disaster risks, ultimately improving response efforts.
4	Fadilah[5] (2022)	Application of the K-Means Algorithm Method for Clustering Landslide Prone Areas in Central Java Province	This research clusters landslide-prone areas based on event frequency, using DBI to determine the optimal number of clusters (k=3), providing valuable input for BPBD's future policy decisions, with special focus on highly vulnerable areas.
5	Dwitiyanti et al[4] (2023)	Implementation of K-Means Method in Classterization of Earthquake Prone Areas in Indonesia	This case study reviews the grouping of earthquake-prone areas in Indonesia using K-means and the latest data. The optimal number of clusters is k = 6, based on a Silhouette index of 0.3245, with one cluster having 5,125 earthquake events and another 632.
6	Setiawan et al[17] (2022)	Clustering of Areas Vulnerable to Disasters in the Form of Ground Movements and Earthquakes in Indonesia	This study explores spatial clustering techniques, including DBSCAN, Common Nearest Neighbor (CNN) and K-Medoids, for grouping areas vulnerable to ground movement and earthquakes in Indonesia. It concludes that K-Medoids is the most suitable method based on its silhouette coefficient score for identifying earthquake-vulnerable regions.
7	Purwayoga et al[14] (2023)	Application of Data Mining for Mapping Disaster Prone Areas as a Disaster Preparedness Effort	The study applies Within Cluster Sum of Squares (WSS) to repeatedly assess the grouping quality, finding the best WSS value at 89.8% with 5 clusters. It also produces a disaster classification map, where each cluster exhibits distinct disaster types and characteristics.
8	Pratama et al[13] (2022)	K-Means Clustering Algorithm Applied to Group Provinces by the Number of Subdistricts Involved in Disaster Anticipation /Mitigation Efforts	The research shows that 14.71% of provinces have a high level of damming efforts, while 85.29% have a low level. These insights from K-means clustering can guide flood management strategies and policies.
9	Wathoriq and Subandi[20] (2023)	Implementation of the K-Means Clustering Algorithm for Grouping Flood Prone Areas	This research analyzes flood impacts in flood-prone areas using the K-means clustering algorithm, aiming to identify potential vulnerabilities. The results indicate that K-means effectively groups floodplains based on their impact levels.
10	Khoirunnisa and Rahmawati[8] (2024)	Comparison of 2 Cluster Methods for Grouping Disaster Intensity in Indonesia	A comparison of K-Means and K-Medoids was conducted to identify patterns and evaluate accuracy. The results show that K-Means had a lower DBI value (0.425) compared to K-Medoids (0.939), indicating better clustering performance.

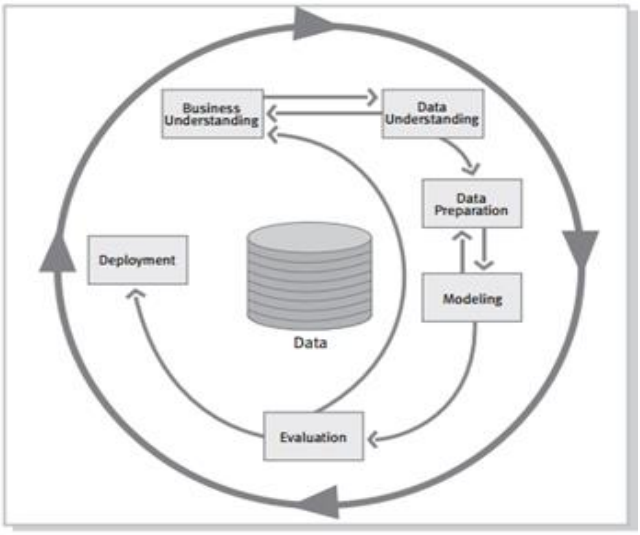

Fig. 1: Phase Crisp-DM

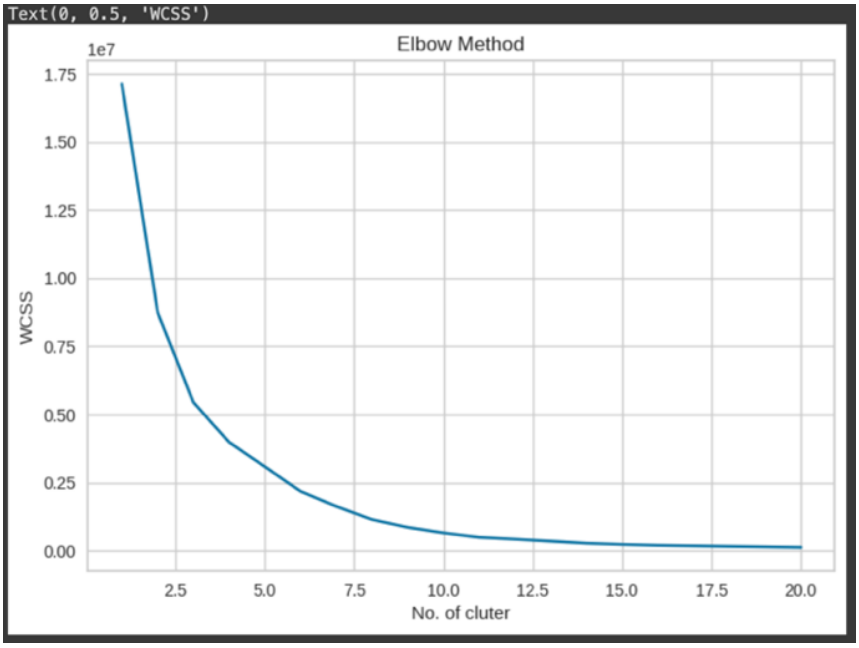

Fig. 2: Graphic cluster elbow method

Table 2
Standard Value Silhouette Coefficient.

Value	Category
0.71 – 1.00	Strong Structure
0.51 – 0.70	Medium Structure
0.26 – 0.50	Weak Structure
≤ 0.25	No Structure

Note that several techniques can be used for the same data mining problem. In addition, if necessary, the process can return to the data processing stage in a format that meets the specific needs of a particular data mining technique[11].

Evaluation: During the evaluation phase, the quality and effectiveness of one or more models are assessed before they can be used. First, it is important to determine whether the model can achieve the desired goals. Then, identify any important business or research questions that still need to be addressed. Finally, decide how the results of the model will be used[11].

To measure the accuracy and quality of the clusters obtained during the modeling stage, researchers use the Silhouette coefficient method. This helps the K-Means algorithm to select the optimal number of clusters. The silhouette coefficient value ranges from -1 to 1. If the value is close to 1, it indicates that the object is in the correct cluster and is quite far from other clusters. However, if it is close to -1, the object is placed in the wrong cluster. Table 2 reflects the accuracy level of the Silhouette coefficient measurement. This evaluation stage evaluates whether the modeling applied is appropriate and suitable for this research case and whether it achieves the desired objectives. Based on the evaluation results, decide whether to continue to the next step or start over if the goal is not achieved.

Deployment: In this phase, we attempted to reveal the results of applying data mining using the grouping method with the implemented K-Means algorithm. In contrast to prior studies, we visualized the clustering results on a digital map using QGIS software.

Results and Discussion

Business understanding: In the initial stage, our aim is to comprehend one's business needs and clarify the data mining challenge to facilitate the achievement of objectives. For this research, we are focused on comprehending the issue of disasters in Indonesia. These are unforeseen occurrences that can pose a threat to the well-being and livelihood of individuals. They are triggered by either natural or human factors that influence the environment and society. The impact of such events on humanity can be severe, even resulting in fatalities.

Data understanding: The dataset used in this research is Indonesian disaster data in 2023. The data source comes from the Statistics Indonesia (BPS). The disaster dataset consists of 34 datasets including province, year, landslides, floods, flash floods, earthquakes, tsunamis, extreme waves, extreme weather, volcanic eruptions, forest fires and droughts.

Data preparation: Data preparation is a step taken to prepare a dataset to suit the needs of the modeling stage. This stage has three steps: data selection, data pre-processing and data transformation. The disaster dataset from 2023, a prime example of our meticulous data selection process, contains 34 dates. From a total of 12 attributes, we carefully selected only 10, namely landslides, floods, flash floods, earthquakes, tsunamis, extreme waves, extreme weather, volcanic eruptions, forest fires and droughts.

Modeling: The next phase is the modeling stage, we use K-means to determine the number of clusters used in the data grouping process. They also used the elbow method. The results are shown in figure 2. From the results of the Elbow method, it can be concluded that the data that has been processed, produces four optimal clusters. The optimal clusters in this method are the points that form Elbow method.

Evaluation: At this stage, the researcher uses the Silhouette coefficient technique to measure or test the quality of the clusters previously obtained at the modeling stage. Researchers measure and test 2 to 5 clusters. Table 3 shows the accuracy results for each cluster. Table 3 shows that cluster 2 obtained the highest Silhouette value (0.62) from the K-means process compared to clusters 3, 4, 5, or 6. Based on table 2, cluster 2 is included in the medium structure criteria. The Silhouette coefficient above 0.51 results indicates that the cluster quality test for grouping disasters in Indonesia in 2023 using the K-means algorithm shows very moderate quality for cluster 2.

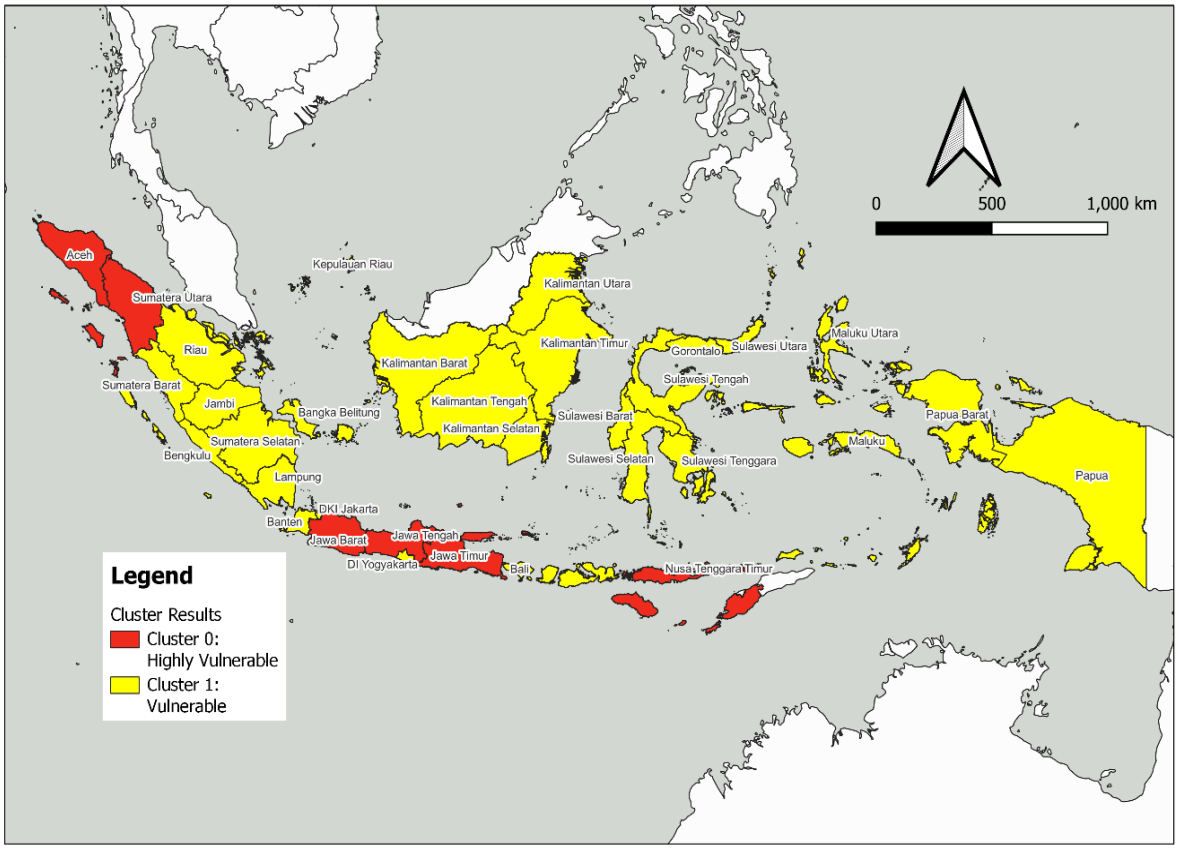

Fig. 3: Clustering of disaster areas in Indonesia into two distinct groups

Table 3
Standard Value Silhouette Coefficient.

S.N.	Cluster	Silhouette Coefficient Accuracy
1	2	0.62
2	3	0.60
3	4	0.45
4	5	0.44
5	6	0.43

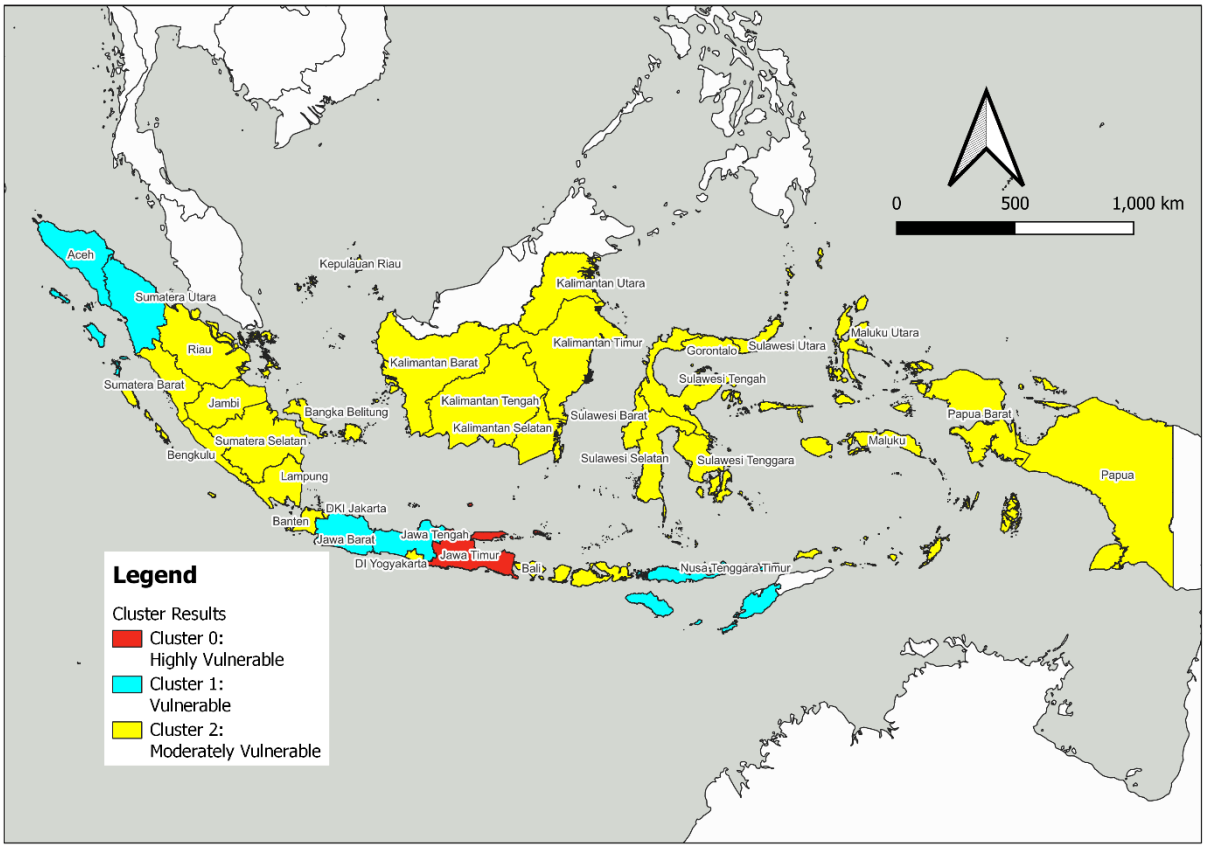

Fig. 4: Clustering of disaster areas in Indonesia into three distinct groups

Deployment: Based on the clustering process with the K-Means algorithm using the Google Colabs tool with the Python programming language, it is suitable for grouping disaster areas in Indonesia because it has the highest Silhouette coefficient value. Furthermore, utilizing geographic information systems for data visualization offers a distinct advantage in illustrating the spatial distribution of disaster intensity. The cluster category is divided into cluster 0 which receives 28 pieces of data and cluster 1, which receives six pieces of data. An excellent way to use the K-Means algorithm is to classify disasters in Indonesia. This is proven to produce a Silhouette coefficient value of 0.62. Details of the data generated for each cluster formed can be seen in figure 3.

After conducting research and testing on clustering data related to disasters that occurred in Indonesia in 2023, researchers found that cluster 0 was included in the power group of areas very prone to disasters, totaling six provinces in Indonesia, while the other clusters were areas at risk of being vulnerable to disasters, with twenty-eight provinces as members. As part of the additional analysis, the study produced regional vulnerability maps using three cluster (Figure 4) and four cluster (Figure 5). These results highlight the differences in vulnerability levels across provinces, enabling comparison with cluster 2 outcomes for informed decision-making.

The clustering results reveal a significant differentiation between regions with high and low disaster intensity. These distinctions highlight the urgent need for tailored disaster mitigation strategies, particularly in high-risk areas, to effectively reduce the potential loss of life and property. Provinces categorized in the high-intensity cluster are characterized by a greater frequency and severity of disasters, often compounded by factors such as higher population density and vulnerability due to inadequate infrastructure. The combination of risk elements from multiple hazards leads to increased exposure and heightened risk. On the other hand, areas with lower-intensity clusters

face fewer disasters. Nonetheless, this does not imply that disaster management planning should be neglected in development efforts. Proactive mitigation and preparedness strategies are crucial in minimizing the effects of extreme events when they happen, ultimately supporting the goal of achieving sustainable resilience.

The clustering analysis conducted in this study directly supports the research objective of providing actionable recommendations for disaster mitigation in Indonesia. By identifying provinces in high-intensity disaster clusters, this study not only highlights regions that require immediate attention but also informs the development of targeted mitigation strategies. These findings suggest that BNPB and local Governments should prioritize resource allocation and disaster preparedness initiatives in these high-risk areas, ensuring that the most vulnerable populations are better protected against the adverse impacts of disasters.

Future studies need to analyze disaster intensity on a smaller regional scale. A micro-zonation approach will enable the identification of areas with more detailed risk levels. In addition, incorporating historical disaster data over a longer time period will be useful for identifying patterns and trends in disaster occurrences. This will facilitate local Governments in allocating resources to develop proactive strategies to reduce the consequences of extreme events in the future.

Conclusion

This study employed the K-means clustering algorithm to classify Indonesian provinces based on disaster intensity in 2023. The analysis identified two primary clusters—28 provinces with high disaster vulnerability and 6 with moderate intensity. Additionally, further analysis revealed the potential for grouping into three and four clusters, offering deeper insights into regional disaster risks.

These findings underscore the importance of focusing disaster mitigation efforts on the most vulnerable regions, guided by the detailed clustering results. The study also demonstrates the utility of data mining in disaster management, suggesting similar approaches could enhance other aspects of disaster risk reduction. In conclusion, the K-means algorithm not only enhances the understanding of disaster distribution across Indonesia but also supports more effective disaster preparedness strategies, contributing to safer and more resilient communities.

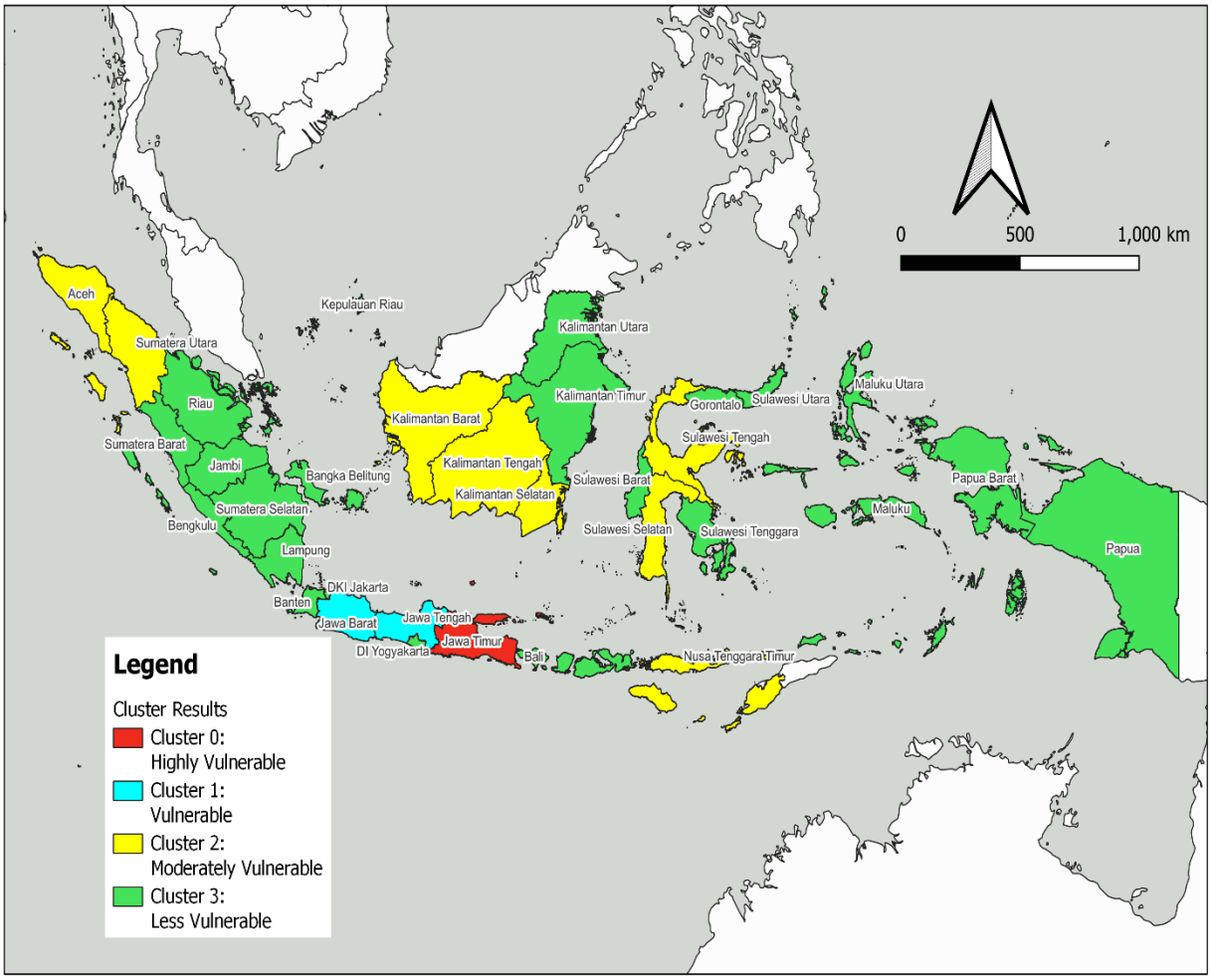

Fig. 5: Clustering of disaster areas in Indonesia into four distinct groups

Acknowledgement

The authors would like to thank the facilities and funding from Universitas Budi Luhur, Jakarta, Indonesia for the completion of this research.

References

1. Aditya A., Jovian I. and Sari B.N., Implementation of K-Means clustering for the national junior high school examination in Indonesia 2018/2019, *Jurnal Media Informatika Budidarma*, **4(1)**, 51-58 **(2020)**

2. Amaliah R., Tohidi E., Wahyudin E. and Rizki Rinaldi A., Grouping disaster data by region using the K-Means algorithm, *Jurnal Mahasiswa Teknik Informatika*, **7(6)**, 35-51 **(2023)**

3. Brzozowska J., Pizon J., Baytikenova G., Gola A., Zakimova A. and Piotrowska K., Data engineering in CRISP-DM process production data - case study, *Applied Computer Science,* **19(3),** 83–95 **(2023)**

4. Dwitiyanti N., Ayu Kumala S. and Dwi Handayani S., Implementation of K-Means method in classterization of earthquake prone areas in Indonesia, *Prosiding Seminar Nasional UNIMUS*, **6**, 1029-1037 **(2023)**

5. Fadilah N., Application of the K-Means algorithm method for clustering landslide prone areas in Central Java Province, *Jurnal BATIRSI*, **6 (1)**, 1-5 **(2022)**

6. Firdaus H., Cluster analysis using K-Means and Fuzzy C-Means in grouping provinces according to data on the intensity of disasters in Indonesia for 2017-2021, *Junal Ilmiah Matematika*, **10(1)**, 50-60 **(2022)**

7. Fransiska N., Anggraeni D. and Enri U., Grouping poverty data for West Java Province using the K-Means algorithm with silhouette coefficient, *Jurnal Teknologi Informasi Komunikasi*, **9**, 29–35 **(2022)**

8. Khoirunnisa F. and Rahmawati Y., Comparison of 2 cluster methods for grouping disaster intensity in Indonesia, *Jurnal Informatika Dan Teknik Elektro Terapan*, **12(1)**, 68-79 **(2024)**

9. Mahroji D., Bernanthos B. and Ratnasih C., Estimating the macroeconomic impact of disasters in Indonesia, Proceedings of the 3rd International Conference on Law, Social Science, Economics and Education (ICLSSEE), 1-9 **(2023)**

10. Mariam S., Handayani F. and Jualiane C., Application of the K-Means clustering algorithm to determine priorities for recipients of home assistance due to disasters, *Jurnal Teknik Informatika Dan Sistem Informasi*, **10(1)**, 231-240 **(2023)**

11. Novita R., Nur Khomarudin A., Aulia R., Yuditihwa A. and Ayuri A., Application of the K-Means algorithm and its analysis to determine student study completion strategies, *Agustus*, **22(2),** 401–413 **(2023)**

12. Nur Khomarudin A., Zakir S., Novita R., Endrawati, Mat M.Z.B.A. and Maiyana E., K-Mean Clustering Algorithm in Grouping Prospective Scholarship Recipients, *Journal of Physics*, Conference Series, **1779(1)**, 012007 **(2021)**

13. Pratama Y., Hendrawan H., Rasywir E., Carenina B.T. and Anggraini D.R., Application of the K-Means clustering algorithm to group provinces based on the number of villages/subdistricts with efforts to anticipate/mitigate disasters, *Building of Informatics, Technology and Science (BITS)*, **4(3)**, 1232-1240 **(2022)**

14. Purwayoga V., Astra Mikail A., Faridah S.D.N. and Retnowati V.A., Application of data mining for mapping disaster prone areas as a disaster preparedness effort, *Jurnal Teknoinfo*, **17(1)**, 319-327 **(2023)**

15. Rohman D., Annisa R., Efendi D.I. and Solahudin D., Disaster clustering using K-Means in the West Java Region, *Jurnal Mahasiswa Teknik Informatika,* **8(1)**, 493-500 **(2024)**

16. Rohman N. and Wibowo A., Clustering Of popular spotify songs in 2023 using K-Means method and silhouette coefficient, *Jurnal Pilar Nusa Mandiri*, **20(1),** 18–24 **(2024)**

17. Setiawan I.N., Krismawati D., Pramana S. and Tanur E., Clustering of areas vulnerable to disasters in the form of ground movements and earthquakes in Indonesia, Prosiding Seminar Nasional Official Statistics 2022, **1**, 669-676 **(2022)**

18. Wahidah N., Juwita O. and Nurman Arifin F., Grouping disaster prone areas in Jember Regency using the K-Means clustering method, *Informatics Journal*, **8(1)**, 22-29 **(2023)**

19. Waluyo M.R.E., Saputra P.Y. and Dien H.E., Clustering landslide areas based on regional and geographical impacts using the K-Means method (Case study: Districts and cities in East Java), Prosiding Seminar Informatika Aplikatif Polinema, **6(1)**, 83-91 **(2020)**

20. Wathoriq T. and Subandi, Implementation of the K-Means clustering algorithm for grouping flood prone areas, Prosiding Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI), **2(2),** 153-159 **(2023**)

21. Yulianto T., Faizzatur Rahmah A. and Amalia R., Clustering disaster areas in Indonesia using the Fuzzy C-Means method, *Jurnal UJMC*, **9(2)**, 29–39 **(2023).**

(Received 01st September 2024, accepted 08th October 2024)