K-Medoids and Support Vector Machine in Predicting the Level of Building Damage in Earthquake Insurance Modeling

Destriana Aulia Rifaldi[1] | *Universitas Islam Indonesia, Yogyakarta, Indonesia*
Atina Ahdika[2] | *Universitas Islam Indonesia, Yogyakarta, Indonesia*

Received 13.3.2024, **Accepted** (reviewed) 25.3.2024, **Published** 13.9.2024

Abstract

Yogyakarta, an Indonesian province prone to earthquakes, frequently suffers extensive damage to buildings, necessitating insurance coverage to mitigate potential losses. This study aims to forecast earthquake insurance premiums by predicting building damage levels resulting from earthquakes. Utilizing data from buildings affected by the June 30, 2023, earthquake in Yogyakarta, we employ K-Medoids Clustering and Support Vector Machine (SVM) to predict two categories of building damage: minor (labelled as 1) and heavy (labelled as 2). The total premiums for minor damage range from approximately USD 86.55 to USD 288.50, while for heavy damage, they range from USD 120.05 to USD 400.18 using the K-Medoids algorithm. Meanwhile, premiums for minor damage range from USD 83.14 to USD 277.13, and for heavy damage, they range from USD 223.67 to USD 745.55 using the SVM algorithm.

Keywords	***DOI***	***JEL code***
Clustering, disaster, earthquake, Yogyakarta, insurance, premium	*https://doi.org/10.54694/stat.2024.13*	*C10, C38, C55, G22*

INTRODUCTION

Disasters are events that occur suddenly and unpredictably resulting in great losses to human life and the environment. Disasters can occur naturally or because of human activity (Makwana, 2019). Disasters that occur naturally include tornadoes, landslides, earthquakes, tsunamis, and erupting mountains. While disasters caused by humans include floods, pollution, and leakage of factory waste. According to McFarlane et al. (2006), disasters are phenomena that can cause trauma to individuals, starting from critical and time-limited conditions and occurring because of nature, technology, and even humans. Indonesia is not spared from natural disasters that threaten such as earthquakes. There are many reasons why Indonesia often experiences earthquakes. According to the United States Geological Survey (USGS),

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Yogyakarta, 55584, Indonesia. E-mail: 20611148@students.uii.ac.id, phone: (+62)895324624841.
[2] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Yogyakarta, 55584, Indonesia. E-mail: atina.a@uii.ac.id, phone: (+62)8172384386. ORCID: <https://orcid.org/0000-0001-9161-227X>.

most earthquakes and volcanic eruptions do not occur randomly, but occur in certain regions, namely the Pacific ring of fire. This area is a confluence between the Pacific plate and various tectonic plates around it, so this area is a very seismically and volcanically active. The fact that Indonesia has a complex and active tectonic arrangement and has a geographical location at the confluence of four tectonic plates supports that Indonesia will not be separated from earthquakes. These tectonic plates include the Pacific Plate, Eurasian Plate, Philippine Sea Plate, and Indo-Australian Plate (National Earthquake Study Center (Indonesia), & Research and Development Center for Housing and Settlements (Indonesia), 2021).

According to Meteorology Climatology and Geophysics Council of Indonesia (BMKG) in 2022, data states that within three years earthquakes in Indonesia have increased. A total of 8 264 earthquakes in 2020, increased in 2021 to 10 519 earthquakes and continued to increase until 2022 to 10 792 earthquakes. The location of this earthquake spread across all provinces in Indonesia from Sabang to Merauke. Of all provinces in Indonesia, Yogyakarta is one of the provinces with a high frequency of earthquakes. In 2006, precisely on May 26, a large earthquake measuring 6.3 on the Richter scale shook Yogyakarta which had a depth of 11.3 km. This earthquake resulted in a lot of losses and casualties. Recently, there has been another major earthquake in Yogyakarta on June 30, 2023. According to the report of the Head of the Regional Disaster Management Agency (BPBD) of the Special Region of Yogyakarta, a tectonic earthquake occurred at 19.57 Indonesian Time with the center located in the South Indian Ocean of Java with a depth of 67 km. All kinds of natural disasters certainly cause losses both material and non-material. This study specifically discusses the earthquake that occurred in Yogyakarta on June 30, 2023, and its impact, especially on buildings. Based on this information, one way to minimize the risk of loss from building damage caused by an earthquake is the use of insurance. Referring to Article 246 of the Commercial Law Code (KUHD) states that insurance is a contract in which the insurer promises to the insured through the collection of premiums, to compensate for losses, damages, and even loss of profits caused by an uncertain event (Santri, 2017). One part of insurance is loss insurance which contains disaster insurance. At present, seismic risk is a pressing issue for both public authorities and private organizations due to the potential for numerous fatalities and considerable economic losses resulting from a seismic event (Hofer et al., 2022).From 2000 to 2016, the average direct economic loss in the form of damage to buildings and supporting objects caused by natural disasters that occurred in Indonesia reached around IDR 22.8×10^12 or USD 1 414 585 096.80 and the possibility of losses due to natural disasters will increase in the future if efforts to reduce, prepare, and transfer risks are not carried out (Fiscal Policy Agency, Ministry of Finance of the Republic of Indonesia, 2018). Risk reduction, setup, and transfer efforts can be run through insurance. Therefore, a good disaster insurance model, especially earthquake insurance is needed in Indonesia.

The purpose of this research is to provide an overview of the calculation of earthquake insurance premiums that result in material losses, especially for affected buildings, by employing k-medoids and Support Vector Machine in determining the level of building damage. In this study, we simulate the calculation of premiums that must be paid by customers to the insurer regarding loss insurance from earthquakes based on actual data from the Regional Disaster Management Agency of Bantul, Yogyakarta, which includes variables such as impact, urban village, latitude, longitude, damage, and Peak Ground Acceleration (PGA). Premium calculation simulation is carried out by predicting the level of damage to buildings due to earthquakes in advance. There are many statistical methods that can be used in predicting data, one of which is k-medoids and *Support vector machine* (SVM). K-medoids is the process of grouping data into certain classes that have the same characteristics, while SVM is one of the statistical methods that is usually used to classify and predict by finding hyperplane or a delimiter to separate two sets of data from two different classes (Octaviani et al., 2014). The k-medoids and SVM methods have differences. In the k-medoids method, there is no need for class labels, but in SVM there are class labels that are used to build models in prediction process. This study compares the two methods for classification

of the level of building damage. We then simulate the calculation of premiums caused by the level of damage resulting from each method.

1 REVIEW OF LITERATURE

The 2018 study by Agustian Noor titled "Comparison of Ordinary Support Vector Machine and Particle Swarm Optimization-Based Support Vector Machine Algorithms for Earthquake Prediction" utilized earthquake data from South Sumatra spanning from 2014 to 2020. The research aimed to analyze earthquake occurrences in North Sumatra by comparing the SVM and SVM-PSO methods, with their performance measured using Root Mean Square Error (RMSE). The results of this study showed that the RMSE value of SVM was 9.720, which was lower compared to the RMSE value of SVM-PSO, which was 37.685 (Noor, 2018).

The 2018 study by Devni Prima Sari et al. titled "Application of Bayesian Network Model in Determining the Risk of Building Damage Caused by Earthquakes" utilized variables from data on building damage caused by the 2009 earthquake in West Sumatra. The data included three independent variables: building structure, PGA (Peak Ground Acceleration), and soil type, as well as one dependent variable: the level of building damage. This research aimed to minimize potential building losses due to earthquakes by predicting the likelihood of building damage at a specific location using a Bayesian network model. The results of this study indicated a 33% probability for light and severe building damage and a 34% probability for moderate building damage, with an accuracy rate of 66% (Wibowo & Institute of Electrical and Electronics Engineers, n.d.).

Previous research in 2019 by Devni Prima Sari et al., titled "K-means and Bayesian Networks to Determine Building Damage Levels," used 7 variables consisting of 4 dependent variables including construction, landslide risk, PGA, and damage, and 3 independent variables including close to faults, slope, and epicenter distance from building unit data of the 2009 West Sumatra earthquake. This research aimed to determine the level of building damage due to earthquakes. The results of this study showed that the levels of light, moderate, and high building damage were 35.46%, 35.14%, and 29.4%, respectively, with an accuracy rate of 70% (Sari et al., 2019).

Research in 2019 by Mariana Yusoff et al., titled "Hybrid backpropagation neural network-particle swarm optimization for seismic damage building prediction," used data obtained from IDARC-2D software from 35 buildings across Malaysia, including 1-story to 35-story buildings. This study used 7 variables including age, number of bays, height, length of seismic zone, natural period, ground acceleration, and building damage index. This research aimed to predict earthquake damage to buildings using hybrid backpropagation neural network and particle swarm optimization (BPNN-PSO). The results of this study showed that BPNN-PSO demonstrated better results with an accuracy of 89% compared to backpropagation neural network with only 84% (Yusoff et al., 2019).

2 METHODS

Cluster analysis or group analysis is one of the statistical research methods that help us to easily classify a set of objects based on information from data into different small clusters. However, objects in each cluster have an affinity or similar characteristics. The groups formed have high internal homogeneity and high external heterogeneity. It can also be interpreted that group analysis maximizes distance between objects and minimizes similarities between groups. Each object is classified according to the distance or proximity of objects to one another, while each variable is classified according to the size of its correlation (Harnanto et al., 2017). This means that grouping is done based on the proximity of the distance between data and the correlation between data variables. High variable correlation allows data to be of the same class. But the absence of correlation between variables will increase the results of grouping. The definition of cluster in data mining is a grouping of several data or objects from clusters (groups) so that each cluster

contains information that is as similar as possible and different from objects in other clusters. In general, clustering methods are divided into two types, namely hierarchical and non-hierarchical ones (Sahriman et al., 2019). Clustering algorithms aim to identify groups of objects that are similar based on their attribute values (Harikumar and Surya, 2015). This study used K-Medoids or can be called Partitioning Around Medoids (PAM), that is a method of grouping n partition objects into k clusters. This grouping uses an object in a set of objects that can represent the cluster. These objects are called medoids which are centrally located in a cluster that is formed. Cluster formation is done by calculating the proximity between medoids objects and non-medoids objects (Musfiani, 2019). K-medoids is a clustering algorithm like k-means, but it is more tolerant to outliers. The main idea of k-medoids is to identify the center of a cluster using k clusters generated randomly (Mohemad et al., 2022). The algorithm of k-medoids is as follows (Mohemad et al., 2022):

1. Determine the number of clusters.
2. Determine k cluster centers randomly.
3. Calculate the distance of each object to the nearest cluster using the gower distance method,

$$d(x_i, x_j) = \frac{\sum_{k=1}^{p} w_k \cdot d_k(x_i, x_j)}{\sum_{k=1}^{p} w_k}, \tag{1}$$

where d(x_i, x_j) is the distance between two objects x_i and x_j, p is the number of variables, w_k is the weight for each variable k, d_k (x_i, x_j) is the distance between two objects x_i dan x_j for variable k.

4. Randomly select non medoid objects in each cluster as new medoid members.
5. Calculate the distance of each non medoid object to the new medoids and assign each non medoid object to the nearest medoid member, then calculate the total distance.
6. Calculate the total deviation S, if the new total deviation is less than the old total deviation, change the position of the new medoid, then make it the new medoid.
7. Repeat steps 4–6 until the medoid does not change (Hermansyah et al., 2024).

After k-medoids are clustered, the results are used as labels in Support Vector Machine (SVM) analysis. SVM is one of the learning methods in machine learning (Wang et al., 2024). Machine learning utilizes past data to build models for predicting future data. Learning, an essential component of artificial intelligence, encompasses diverse statistical, probabilistic, and optimization techniques like logistic regression, artificial neural networks (ANN), K-nearest neighbor (KNN), decision trees (DT), and Naive Bayes (Huang et al., 2018). SVM known for their computational power in supervised learning, are extensively employed in addressing classification, clustering, and regression tasks (Nayak et al., 2015). The Support Vector Machine has demonstrated its effectiveness as a powerful tool for supervised classification (Wang et al., 2024; Huang et al., 2018; Nayak et al., 2015). Vapnik in 1998 introduced the method SVM as one of the methods for classification which basically works in finding the boundary between two classes with the maximum distance of the best closest data through the formation of hyperplane (limit). This limit is obtained by measuring the margin hyperplane and looking for the maximum point. The margin represents the distance between the closest point of each class and hyperplane. This point is commonly referred to as SVM (Achmad Rizal et al., 2019). SVM can handle both linear and nonlinear data. In linear data, hyperplane is easy to find, whereas in nonlinear data, data is simulated into three dimensions first using a function called a kernel. The kernel used is a function used in grouping low-dimensional data into high-dimensional (Mase et al., 2018). Some kernels that can be used are as follows:

- Kernel linear:

$$K(x, y) = x \cdot y, \tag{2}$$

where x is training data and y is testing data.

- Kernel polynomial:

$$K(x, y) = (x \cdot y + c)^d, \tag{3}$$

where x is the training data and y is the testing data. While d is the degree of polynomial.

- Kernel Gaussian RBF (Radial Basis Function):

$$K(x, y) = exp\left(\frac{-||x - y||^2}{2 \cdot \sigma^2}\right), \tag{4}$$

where x is training data and y is testing data.

- Kernel sigmoid:

$$K(x, y) = \tanh(\sigma(x \cdot y) + c), \tag{5}$$

where x is training data and y is testing data. While c is a coefficient.

- Inverse multiquadric kernel:

$$K(x, y) = \frac{1}{\sqrt{||x - y||^2 + c^2}}, \tag{6}$$

where x is training data and y is testing data. While c is a coefficient.

The result of SVM is in the form of a confusion matrix. Confusion matrix contains prediction data and actual data to illustrate how the actual classes and their predictions differ.

Table 1 Confusion matrix

		Actual Value	
		Positive	**Negative**
Predicted Value	Positive	TP	FP
	Negative	FN	TN

Source: Mase et al. (2018)

There are 4 terms as a representation of the results of the classification process in the confusion matrix. True Positive (TP) is true data predicted positive, True Negative (TN) is true data predicted negative, False Positives (FP) is data that is incorrectly predicted to be positive, and False Negative (FN) is data that is incorrectly predicted to be negative.

Based on the results obtained from the confusion matrix, then calculated each value using accuracy, precision, recall, and F-1 score.

a. Accuracy is the value of the accuracy of the model in the classification.

$$accuracy = \left(\frac{(BP + BN)}{(BP + SP + BN + SN)}\right) X\ 100\%\,. \tag{7}$$

b. Precision is the accuracy between data and prediction results.

$$precision = \left(\frac{(BP)}{(BP + SP)} \right) X\ 100\% . \tag{8}$$

c. Recall is a value that indicates the success of the model in finding information.

$$recall = \left(\frac{(BP)}{(BP + SN)} \right) X\ 100\% . \tag{9}$$

d. F-1 score is a comparison between *precision and recall values*.

$$f1 - score = \left(\frac{2 \times precision \times recall}{precision + recall} \right) X\ 100\% . \tag{10}$$

After applying both methods in classifying and predicting the level of building damage, the results will be used to simulate premium calculations in earthquake disaster insurance following these steps (Yucemen, 2005).

- Defining the probability of damage for each level of building damage $i(P_i(DB))$:

$$P_i(DB) = \frac{N_i(DB)}{N(DB)}, \tag{11}$$

with $N_i(DB)$ is the number of damaged buildings in the earthquake area with type i = 1, 2, 3, where 1, 2, and 3 represent minor, moderate, and heavy damage, respectively, and $N(DB)$ is the total number of buildings that suffered damage caused by earthquake.

- Calculate the mean damage ratio for each level of building damage $i(MDR_i)$:

$$MDR_i(M) = \sum_{DB} P_i(DB) \times CDR_{DB} , \tag{12}$$

where CDR_{DB} is the corresponding central damage ratio or ratio of the number of damaged buildings and the overall buildings in the earthquake area.

- Calculating the expected annual damage ratio for level of building damage $i(EADR_i)$:

$$EADR_i = \sum_M MDR_i(M) \times AP_M , \tag{13}$$

where $MDR_i(M)$ is the mean damage ratio for the level of building damage i that experienced an earthquake with intensity M and AP_M is the annual probability of an earthquake with intensity M occurs in an area.

- Calculating the pure risk premium for level of building damage $i(PRP_i)$:

$$PRP_i = EADR_i \times BIV , \tag{14}$$

then the pure risk premium can be calculated based on the building insured value (BIV).

- Calculating the total earthquake insurance premium for level of building damage $i(TP_i)$:

$$TP_i = \frac{PRP_I}{1 - LF}, \tag{15}$$

where *LF* is load factor which is defined as hidden uncertainties such as administrative expenses, business taxation, and benefits for the Insurance Company.

3 RESULTS AND DISCUSSION

3.1 Clustering

The solution for clustering analysis using R Studio generally uses the PAM method and there will be three steps taken, namely determining the distance between observations using gower distance because the data used are numerical and categorical mixed data, determining the number of clusters, and clustering. The gower distance will compare the data pair on a scale of 0 to 1. If the two data compared are close to 0 then the data are close together. Conversely, if the two data compared are close to 1 then the data are far apart. In R Studio, the calculation of gower distance is contained in cluster packages with the daisy() function and in Table 2 and 3 are examples of the closest and most distant data.

Table 2 The most nearby data

Data	Impact	Urban village	Latitude	Longitude	Epicenter distance	Damage	PGA
38	House	Srigading	–7.808	110.456	19.91480	Cracked wall	3.268858
37	House	Srigading	–7.807	110.455	19.91475	Cracked wall	3.268874

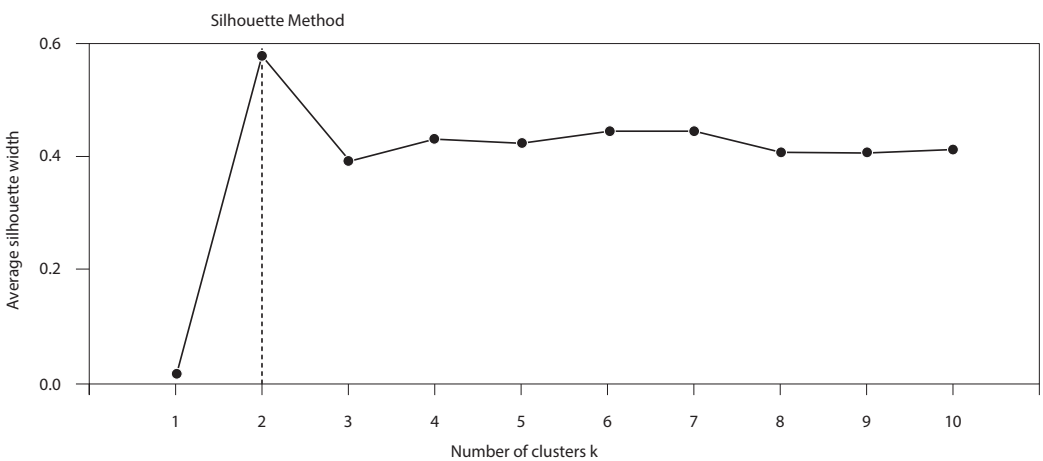
Source: Own elaboration

Table 3 The most distant data

Data	Impact	Urban village	Latitude	Longitude	Epicenter distance	Damage	PGA
69	House	Srigading	–8.005	110.266	19.94268	4-point cracked wall	3.261004
24	Stall	Parangtritis	–6.404	106.817	19.57868	Roof collapsed	3.365682

Source: Own elaboration

Figure 1 Optimal number of clusters



Source: Own elaboration

Based on Table 2 and Table 3, data 37 and 38 are very close together with a minor difference in numerical observations of 0.00005 in the epicenter distance variable compared to data 69 and 24 which are very far apart because there are many differences in each numerical variable, such as the epicenter distance variable which has a difference of 0.364. Furthermore, the determination of number of clusters using the silhouette method with the help of R Studio produces a graph in Figure 1.

Figure 1 above shows that as many as 2 clusters are the best. Then the next step is to group the data into 2 groups with the PAM method using the pam() function in R Studio. This function produces two cluster centers that can be used to divide the data into two clusters.

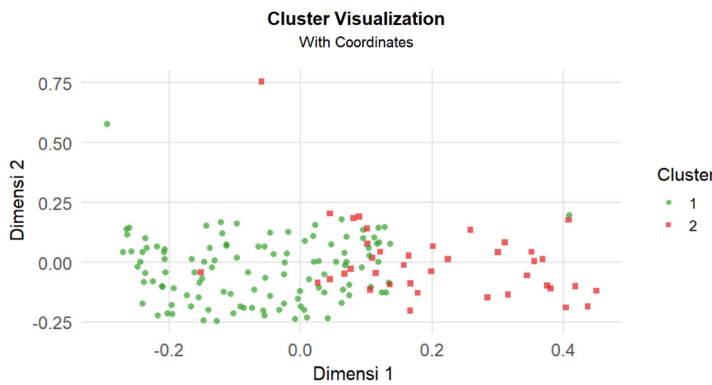
Table 4 Cluster center

Cluster	Data	Impact	Urban village	Latitude	Longitude	Epicenter distance	Damage	PGA
1	41	House	Srigading	–7.811	110.459	19.91497	Cracked walls	3.268811
2	131	Educational facilities	Parangtritis	–7.990	110.316	19.93937	Classroom	3.261934

Source: Own elaboration

Based on the results of the functions that have been executed in Table 4, each cluster center is at data 41 for cluster 1 and data 131 for cluster 2, the results of clustering are listed in the Figure 2.

Figure 2 Cluster results graph



Source: Own elaboration

Based on Figure 2, cluster 1 has 114 data members and cluster 2 has 40 data members. To see which clusters are classified as minor damage and which clusters are classified as heavy damage, profiling is carried out. Profiling is done by calculating the average for numerical data for epicenter and PGA distance variables.

Based on the results in Table 5, cluster 2 has smaller epicenter distances than cluster 1, meaning that the area of earthquake-affected buildings close to the epicenter is the affected building area classified in cluster 2. Therefore, buildings in cluster 1 are minorly affected, while buildings in cluster 2 are heavily affected.

According to the level of damage grouped into 3, namely minor, moderate, and heavy damage. Cluster 1 has a higher percentage of minor damage than cluster 2. While cluster 2 has a higher percentage of moderate and heavy damage than cluster 1 even though the different percentage of level of heavy damage

Table 5 Epicenter and PGA distance variable cluster profiling

	1	2
Epicenter distance	19.9203238	19.9199753
PGA	3.2673206	3.26745224

Source: Own elaboration

Table 6 Percentage of damage from K-Medoids

Type of damage buildings		Cluster 1		Cluster 2	
Minor wall crack	Minor	45.6%	53.5%	12.5%	12.5%
Broken pipelines		0.9%		-	
Roof collapsed		7.0%		-	
Broken tiles and cracked walls	Moderate	25.4%	31.6%	22.5%	72.5%
Sloping and cracking walls		4.4%		-	
Some points of the wall are broken		1.8%		50.0%	
Heavy wall cracks	Heavy	6.1%	14.9%	2.5%	15.0%
Wall collapsed		8.8%		12.5%	

Source: Own calculations

is only 0.1%. Therefore, cluster 1 members can be called earthquake-affected observations with minor damage and cluster 2 members can be called earthquake-affected observations with heavy damage.

3.2 Support vector machine

SVM analysis uses data labels to train its model. Labeling in this case is obtained from the results of clustering k-medoids. The data used was resampling data from the original data to 9 000 from 154 data. Resampling data is conducted to improve the accuracy of SVM. The data uses numerical variables from the original data, namely latitude, longitude, epicenter distance, PGA, and cluster as labeling. SVM is done by dividing data into training and testing data in a ratio of 70:30. The results of the classification and prediction of the level of damage to buildings due to the June 30, 2023, Bantul earthquake are described in Table 7.

Table 7 Confusion matrix SVM

	Data training		Data testing	
	Actual data classes		Actual data classes	
Prediction class	Minor damage rate	Heavy damage rate	Minor damage rate	Heavy damage rate
Minor damage rate	4 621 (74.24%)	1 603 (25.76%)	2 017 (75.74%)	646 (24.26%)
Heavy damage rate	35 (46.67%)	41 (53.33%)	20 (54.05%)	17 (45.95%)

Source: Own calculations

Table 7 shows that in the training data there is 74.24% data that is predicted correctly as data with the level of minor building damage and in the testing data there is 75.74% data that is predicted correctly as data with the level of minor building damage. Furthermore, in the training data there is 53.33% data

that is predicted correctly as data with the level of heavy building damage and in the testing data there is 45.95% data that is predicted correctly as data with the level of heavy building damage. The accuracy obtained in training data is 74% and testing data is 75%, meaning that the model built can predict large parts of the data correctly. Then from the prediction results using SVM, each level of damage is grouped to get a percentage of damage, namely prediction 1 is the buildings that are predicted to have a minor level of damage and prediction 2 is the buildings that have a heavy level of damage. The details of the damage percentages are explained in Table 8.

Table 8 SVM result damage percentage

	Prediction 1	Prediction 2
Minor	47.9%	0.0%
Moderate	39.8%	51.3%
Heavy	12.1%	48.6%

Source: Own calculations

Based on Table 8, SVM predicts that none of the buildings suffered minor damage in prediction class 2.

3.3 Building damage insurance premium calculation simulation

From the results of the classification of building damage levels based on k-medoids and SVM, both can be used as a reference in the calculation of simulated building damage premiums caused by the Bantul earthquake on June 30, 2023, see Table 9.

Table 9 Probability of damage

Probability of damage $P_i(DB)$	K-Medoids		SVM	
	1	2	1	2
Minor damage	0.535	0.125	0.479	0
Moderate damage	0.316	0.725	0.398	0.513
Heavy damage	0.149	0.150	0.121	0.486

Source: Own calculations

Before calculating the average damage ratio, an appropriate central damage ratio (CDR_{DB}) is required and refers to the average damage ratio of the Yogyakarta earthquake in 2006 of 5.9 Scale of Richter (SR) (Arrie and Amin, 2018), see Table 10.

Table 10 Average damage ratio of 2006 Yogyakarta earthquake

Damage rate	Damage ratio of the 2006 Yogyakarta earthquake
Minor damage	0.0121
Moderate damage	0.1399
Heavy damage	0.5790

Source: Arrie and Amin (2018)

Then calculate the average damage ratio ($MDR_i(M)$), see Table 11.

Table 11 Average damage ratio

$MDR_i(M)$	K-Medois		SVM	
	Cluster 1	Cluster 2	1	2
Minor damage	0.0064735	0.0015125	0.0057959	0
Moderate damage	0.0440685	0.1014275	0.0556802	0.0717687
Heavy damage	0.086271	0.08685	0.070059	0.281394
Sum	0.136813	0.18979	0.1315351	0.3531627

Source: Own calculations

Calculating the expected annual damage ratio ($EADR_i$) to AP_M is the annual probability of an earthquake with intensity M occurring in a region. During 2023, there have been 49 earthquakes around Bantul, based on BMKG data. Earthquakes of magnitude 6.4 or more occur only once a year. Then $AP_{6.4SR} = 0.02041$. The resulting $EADR_i$ are given in Table 12.

Table 12 Annual damage ratio

$EADR_i$	K-Medois		SVM	
	Cluster 1	Cluster 2	1	2
	0.00279	0.00387	0.00268	0.00721

Source: Own calculations

Then it can be calculated pure risk premium by Formula (15), with BIV is the value of the building insured, for example IDR 300 000 000 (USD 18 612.96), IDR 500 000 000 (USD 31 021.60), IDR 750 000 000 (USD 46 532.40), and IDR 1 000 000 000 (USD 62 043.21), see Table 13.

Table 13 Pure risk premium

BIV	PRP_i			
	K-Medoids		SVM	
	Cluster 1	Cluster 2	1	2
IDR 300 000 000	IDR 837 000	IDR 1 161 000	IDR 804 000	IDR 2 163 000
(USD 18 612.96)	(USD 51.93)	(USD 72.03)	(USD 49.88)	(USD 134.20)
IDR 500 000 000	IDR 1 395 000	IDR 1 935 000	IDR 1 340 000	IDR 3 605 000
(USD 31 021.60)	(USD 86.55)	(USD 120.05)	(USD 83.14)	(USD 223.67)
IDR 750 000 000	IDR 2 092 500	IDR 2 902 500	IDR 2 010 000	IDR 5 407 500
(USD 46 532.40)	(USD 129.83)	(USD 180.08)	(USD 124.71)	(USD 335.50)
IDR 1 000 000 000	IDR 2 790 000	IDR 3 870 000	IDR 2 680 000	IDR 7 210 000
(USD 62 043.21)	(USD 173.10)	(USD 240.11)	(USD 166.28)	(USD 447.33)

Source: Own calculations

After getting a pure risk premium by assuming that the building values are IDR 300 000 000 (USD 18 612.96), IDR 500 000 000 (USD 31 021.60), IDR 750 000 000 (USD 46 532.40), and IDR 1 000 000 000 (USD 62 043.21), the total premium described in Table 14 can be calculated according to (15). In this work, we assume that the load factor is 0.4.

Table 14 Total premium

BIV	TP_i			
	K-Medoids		SVM	
	Cluster 1	Cluster 2	1	2
IDR 300 000 000 (USD 18 612.96)	IDR 1 395 000 (USD 86.55)	IDR 1 935 000 (USD 120.05)	IDR 1 340 000 (USD 83.14)	IDR 3 605 000 (USD 223.67)
IDR 500 000 000 (USD 31 021.60)	IDR 2 325 000 (USD 144.25)	IDR 3 225 000 (USD 200.09)	IDR 2 233 333 (USD 138.56)	IDR 6 008 333 (USD 372.78)
IDR 750 000 000 (USD 46 532.40)	IDR 3 487 500 (USD 216.38)	IDR 4 837 500 (USD 300.13)	IDR 3 350 000 (USD 207.84)	IDR 9 012 500 (USD 559.16)
IDR 1 000 000 000 (USD 62 043.21)	IDR 4 650 000 (USD 288.50)	IDR 6 450 000 (USD 400.18)	IDR 4 466 667 (USD 277.13)	IDR 12 016 667 (USD 745.55)

Source: Own calculations

Based on Table 14, the total premium for building damage due to the Bantul earthquake on June 30, 2023, for minor to heavy damage from the k-medoids algorithm increases gradually as the *BIV* increased. The total premiums for cluster 2 (heavy-affected building) are higher than those of cluster 1 (minor-affected building). However, it does not have significantly different value ranges. The total building damage premium for predictions using the SVM algorithm also increases gradually as the *BIV* increased. However, for prediction 2 (heavy damage levels), there is a significant increase in premiums, which may be due to the absence of buildings classified or predicted to have minor damage. All building units are classified as moderate and heavy types.

CONCLUSIONS

Using the k-medoids algorithm in clustering and Support Vector Machine (SVM) in prediction produces two types of building damage: minor damage as cluster 1 (prediction 1) and heavy damage as cluster 2 (prediction 2). From these two methods, the amount of earthquake disaster insurance premiums that must be paid can be simulated. For the k-medoids algorithm, the premium amount for minor and heavy damage levels does not differ significantly. This is because both cluster 1 and cluster 2 contain buildings with minor, moderate, and heavy damage levels even though with different percentages. Meanwhile, for the SVM algorithm, the premium amount for minor and heavy damage levels differs significantly. This is because there are no buildings predicted to have minor damage. All buildings are predicted to have moderate and heavy damage levels. What can be further developed in research on this topic is the possibility of conducting simulation calculations for claims and improving the accuracy of the SVM method.

ACKNOWLEDGMENT

In the process of this research, the authors would like to express their gratitude to the supervising lecturer from the Department of Statistics of Universitas Islam Indonesia who has helped, feedback, and guidance, as well as supervision, ensuring that this paper is worthy of publication and is expected to be beneficial to the wider audience.

References

ACHMAD RIZAL, R., SANJAYA GIRSANG, I., APRIYADI PRASETIYO, S. (2019). Face Classification Using Support Vector Machine (SVM) (in Indonesian). *Research and E-Journal of Computer Informatics Management*, 3(2).

ARRIE, M., AMIN, R. (2018). *Ratio of Residential Building Damage. Case Study: Aceh Earthquake on July 2, 2013* (in Indonesian).
FISCAL POLICY AGENCY, MINISTRY OF FINANCE OF THE REPUBLIC OF INDONESIA (2018). *Financing Strategy and Disaster Risk Insurance* (in Indonesian).
HARIKUMAR, S., SURYA, P. V. (2015). K-Medoid Clustering for Heterogeneous DataSets [online]. *Procedia Computer Science*, 70: 226–237. <https://doi.org/10.1016/j.procs.2015.10.077>.
HARNANTO, Y. I., RUSGIYONO, A., WURYANDARI, T. (2017). Application of Ward Cluster Analysis Method to Regencies/Cities in Central Java Based on Contraceptive Use [online] (in Indonesian). *GAUSSIAN Journal*, 6: 528–537. <http://ejournal-s1.undip.ac.id/index.php/gaussian>.
HERMANSYAH, M., AFREYNA FAUZIAH, D., SABILIRRASYAD, I., FAIZ FIRDAUSI, M., WAHID, A. (n.d.). *Comparison of K-Means and K-Medoids Algorithms in Students English Skill Clasterization.* 1(1): 1.
HOFER, L., ZANINI, M. A., FALESCHINI, F., PELLEGRINO, C. (2022). Expected losses vs earthquake magnitude curves, for seismic risk mitigation and for insurance purposes [online]. *Procedia Structural Integrity*, 44: 1824–1831. <https://doi.org/10.1016/j.prostr.2023.01.233>.
HUANG, S., NIANGUANG, C. A. I., PENZUTI PACHECO, P., NARANDES, S., WANG, Y., WAYNE, X. U. (2018). Applications of support vector machine (SVM) learning in cancer genomics [online]. In: *Cancer Genomics and Proteomic*s, International Institute of Anticancer Research, 15(1): 41–51. <https://doi.org/10.21873/cgp.20063>.
MASE, J., FURQON, M. T., RAHAYUDI, B. (2018). Implementation of Support Vector Machine (SVM) [online] (in Indonesian). *Algorithm in Cat Disease Classification*, 2(10). <http://j-ptiik.ub.ac.id>.
MCFARLANE, A. C., NORRIS, F. H., GALEA, S., FRIEDMAN, M. J., WATSON, P. J. (2006). *Definitions and Concepts in Disaster Research.*
MOHEMAD, R., MUHAIT, N. N. M., NOOR, N. M. M., OTHMAN, Z. A. (2022). Performance analysis in text clustering using k-means and k-medoids algorithms for Malay crime documents [online]. *International Journal of Electrical and Computer Engineering*, 12(5): 5014–5026. <https://doi.org/10.11591/ijece.v12i5.pp5014-5026>.
MUSFIANI, M. (2019). Cluster Analysis Using Partition Method on Contraceptive Users in West Kalimantan [online] (in Indonesian). *Bimaster: Scientific Bulletin of Mathematics, Statistics, and Its Applications*, 8(4). <https://doi.org/10.26418/bbimst.v8i4.36584>.
NATIONAL EARTHQUAKE STUDY CENTER, RESEARCH AND DEVELOPMENT CENTER FOR HOUSING AND SETTLEMENTS. (2021). *Map of earthquake sources and hazards in Indonesia in 2017 in Indonesia.*
NAYAK, J., NAIK, B., BEHERA, H. S. (2015). A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges [online]. *International Journal of Database Theory and Application*, 8(1): 169–186. <https://doi.org/10.14257/ijdta.2015.8.1.18>.
NOOR, A. (2018). *Comparison of Regular Support Vector Machine Algorithm and Particle Swarm Optimization-Based Support Vector Machine for Earthquake Prediction* (in Indonesian).
OCTAVIANI, P. A., WILANDARI, Y., ISPRIYANTI, D. (2014). Application of Support Vector Machine (SVM). Classification Method on Elementary School (SD) [online] (in Indonesia). *Accreditation Data in Magelang Regency*, 3(4): 811–820. <http://ejournal-s1.undip.ac.id/index.php/gaussian>.
SAHRIMAN, S., KALONDENG, A., KOERNIAWAN, V. (2019). Statistical Downscaling Modeling with Dummy Variables Based on Hierarchical and Non-Hierarchical Cluster Techniques for Rainfall Estimation [online] (in Indonesian). *Indonesian Journal of Statistics and Its Applications*, 3(3). <http://www.climatexp.knmi.nl>.
SANTRI, S. H. (2017). The Principle of Utmost Good Faith in Insurance Contracts [online] (in Indonesian). *UIR LAW REVIEW*, 1(1): 77. <https://doi.org/10.25299/ulr.2017.1.01.165>.
SARI, D. P., ROSADI, D., EFFENDIE, A. R., DANARDONO. (2019). K-means and bayesian networks to determine building damage levels [online] (in Indonesian). *Telkomnika (Telecommunication Computing Electronics and Control)*, 17(2): 719–727. <https://doi.org/10.12928/TELKOMNIKA.V17I2.11756>.
VALERO-CARRERAS, D., ALCARAZ, J., LANDETE, M. (2023). Comparing two SVM models through different metrics based on the confusion matrix [online]. *Computers and Operations Research*, 152. <https://doi.org/10.1016/j.cor.2022.106131>.
WANG, P., GUO, J., LI, L.-F. (2024). Machine learning model based on non-convex penalized huberized-SVM [online]. *Journal of Electronic Science and Technology*, 100246. <https://doi.org/10.1016/j.jnlest.2024.100246>.
WIBOWO, F. W. (n.d.). *2018 International Conference on Information and Communications Technology (ICOIACT)*, 6–7 March 2018, Institute of Electrical and Electronics Engineers.
YUCEMEN, M. S. (2005). Probabilistic assessment of earthquake insurance rates for Turkey [online]. *Natural Hazards*, 35(2): 291–313. <https://doi.org/10.1007/s11069-004-6485-8>.
YUSOFF, M., DNAJIB, F. M., ISMAIL, R. (2019). Hybrid backpropagation neural network-particle swarm optimization for seismic damage building prediction [online] (in Indonesian). *Indonesian Journal of Electrical Engineering and Computer Science*, 14(1): 360–367. <https://doi.org/10.11591/ijeecs.v14.i1>.