Eksakta: Berkala Ilmiah Bidang MIPA
VOLUME 23 NO 03 2022, pp 231-241
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol23-iss03/304

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

Article

Disaster Mitigation Efforts Using K-Medoids Algorithm and Bayesian Network

Devni Prima Sari[1*], Media Rosha[1], Dedi Rosadi[2]

[1]Department of Mathematics, Faculty of Mathematics and Natural Science (FMIPA), Universitas Negeri Padang, Indonesia
[2]Department of Mathematics, Faculty of Mathematics and Natural Science (FMIPA), Universitas Gadjah Mada, Indonesia

Article Info

Article history :

Received January 25, 2022
Revised February 20, 2022
Accepted March 21, 2022
Published September 30, 2022

Keywords :
Mitigation,
K-medoids,
bayesian networks,
earthquake, damage.

Abstract. Disaster mitigation is a series of efforts to reduce disaster risk. One of the disaster mitigation efforts is the supervision of the implementation of spatial planning. Knowing the level of damage to buildings in a region in the event of a disaster can supervise the implementation of spatial planning. To predict the level of damage to buildings in an area, we can use the Bayesian network Model. Bayesian network is an extension of Naive Bayes. There are several types of Bayesian networks based on the variable type, namely discrete Bayesian network, continuous Bayesian network, and hybrid Bayesian network. A discrete Bayesian network is a Bayesian network model in which all the variables involved are discrete. Therefore, if there is a continuous variable, it is necessary to discretize the variable. In this paper, modifications are made to the algorithm commonly used in the clustering process to be used in the discretization process. The algorithm used is the K-Medoids algorithm, where this algorithm uses existing data as a representative of the cluster center. Then, the Bayesian network model and the K-Medoids algorithm were used to determine the level of damage to buildings due to the earthquake that occurred in West Sumatra in 2009. From 25,000 house damage data used in this study, we obtain an accuracy rate is 95.17%.

Corresponding Author :
Devni Prima Sari
Department of Mathematics, Faculty of Mathematics and Natural Science (FMIPA),
Universitas Negeri Padang, Indonesia
Email : devniprimasari@fmipa.unp.ac.id

1. Introduction

There are numerous barriers and limits to determining the exact moment of a natural disaster. On the other hand, natural disasters frequently cause extensive material and non-material losses. By implementing disaster mitigation, we can lessen the impact of disaster-related losses. Catastrophe mitigation is a set of strategies for reducing disaster risk through physical development, disaster awareness, and disaster capacity building (Article 1 paragraph 6 PP No. 21 of 2008 concerning the Implementation of Disaster Management). Disaster mitigation aims to lessen the impact of catastrophes, particularly on the population, serve as a foundation (guideline) for development planning, and enhance public awareness of how to deal with and reduce disaster impact/risk so that people may live and work securely. Supervising the application of spatial planning is one of the catastrophe mitigation strategies. Knowing the extent of damage to buildings in a disaster area can help maintain the performance of spatial planning.

We can utilize the idea of the opportunity to forecast the extent of damage to buildings in a given location. When it comes to the concept of opportunity, Bayes' Theorem is unavoidable. Bayes ' theorem describes the link between the conditional probability of two events, which has critical applications in statistics. Naive Bayes, Hidden Naive Bayes, and Bayesian networks apply Bayes' theorem principles in the categorization process. The Bayesian network is a Naive Bayes extension. Based on variables, discrete, continuous, and hybrid Bayesian networks are the three forms of Bayesian networks. A discrete Bayesian network is a Bayesian network model in which all of the variables are discrete.

In this paper, the author is interested in predicting the level of damage to buildings using the Discrete Bayesian network. In the Discrete Bayesian network, all variables must be discrete. Therefore, if there is a continuous variable, it is necessary to discretize the variable. Discretization is converting a continuous variable into a discrete variable and creating partitions in the range of values that the variable takes. Then a mapping is made between each interval in the partition and the discrete values of the numbers. Once the discretization is performed, the new variable can be treated as an ordinal. Discretization can be seen as one of the possible data preprocessing techniques. These techniques can significantly improve the overall quality of relationships extracted from the data and the time required for analysis [1]. Discretization can or must be applied before using many statistical models. In fact, there are many models designed primarily for processing categorical data, such as Naive Bayes (NB) [2] and Bayesian network (BN) [3]–[7]. Both models examine the relationships between the variables of interest and allow the coexistence of discrete and continuous variables in the dataset under investigation.

Nevertheless, in the case of BN, the hybrid database enforces constraints on the parent-child relationship between variables. Discrete variables only need discrete parents [8], which can be an unrealistic constraint in many applications. Probabilities need to be estimated for BN and NB, making it challenging to handle continuous variables. To avoid this problem, they are generally assumed to be normally distributed, but this assumption does not always reflect the nature of these variables. Moreover, even if the model can handle continuous variables, the learning process is less efficient and effective [9].

After the discretization process for the new variable is carried out, the classification of the level of damage to the building is carried out using the Bayesian network. The Bayesian network (BN) is a graphical model for expressing information about uncertain domains that are probabilistic. Each node represents a random variable, and each edge reflects the related random variable's conditional probability [10]. Bayesian networks are familiar to be applied in various fields, including mining, finance, health, and disaster mitigation.

In the case in this study, we modified the algorithm commonly used in the clustering process to be used in the discretization process. The algorithm used is the K-Medoids algorithm, where this algorithm uses existing data as a representative of the cluster center. Then, the Bayesian network model and the K-Medoids algorithm were used to determine the level of damage to buildings due to the earthquake in West Sumatra in 2009.

2. Method

2.1. Procedure

The research in this study was conducted by analyzing the theories relevant to the problems discussed based on the literature review. The development carried out is considering the discretization analysis of exogenous and endogenous variables to determine the level of damage to buildings using the K-Medoids method. Then, clustering the level of damage to buildings using the Bayesian network model. Details of the research method can be seen in
Figure **1**.

2.2. K-Medoids Algorithm

K-Medoids is a partition clustering approach that reduces the distance between a cluster's labeled and center points. Each K-Medoids or PAM algorithm cluster is centered on an object (medoid). The K-Medoids approach has the advantage of overcoming the K-Means algorithm's flaw of being susceptible to noise and outliers, which can cause objects with great values to depart from the data distribution. Another benefit is that the clustering process' outcomes are independent of the sequence in which the records are entered. Procedure for the K-Medoids algorithm [1], [11]:

a) Set up k cluster centers (number of clusters)
b) Assign all data (objects) to the nearest cluster using the Euclidean distance measurement formula.
c) Choose one object from each cluster at random as a candidate for a new medoid.
d) Calculate the distance between each object in each cluster using the new candidate medoid.
e) Calculate the total deviation (S) by comparing the new distance's total value to the old distance's total value. For $S < 0$, these objects are combined with cluster data to create a new collection of k medoids.
f) Repeat steps 3 to 5 until there is no medoid change to obtain clusters and their respective cluster members.

Not only can the K-Medoids technique be used to group objects, but it can also be used to discretize continuous variables. The number of features is limited to two due to discretization. The first function is a discretized variable, and the second function is an assumed constant auxiliary function.

2.3. Bayesian Network

Bayesian network are graphs made up of nodes and arcs that indicate interactions between variables. Consider the random vector $\boldsymbol{X} = (X_1, .., X_H)$, defined in the state space $\mathcal{X} = \mathcal{X}_1 \times .. \times \mathcal{X}_H$, where $\mathcal{X}_i$ is the state space for X_i, and $\mathcal{X}_i = \left\{x_i^{(1)}, \dots, x_i^{(K)}\right\}$ for $i = 1, \dots, H$. If the variable X_i is affected by the variable X_j, then $(X_j, X_i) \in E$. For the variable X_i, $\Pi_i = \left\{X_j | (X_j, X_i) \in E\right\}$ where Π_i is the parent set for the variable X_i, which is the set of variables in the model whose value is a direct cause of the value of the variable X_i [12].

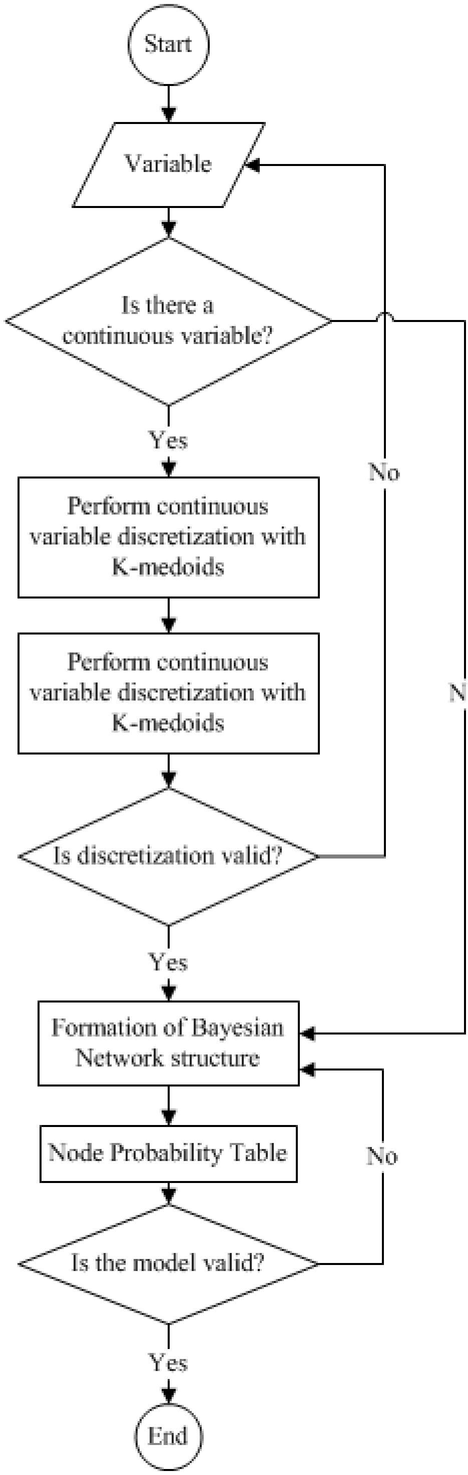

Start
Variable
Is there a continuous variable?
Yes
No
Perform continuous variable discretization with K-medoids
Perform continuous variable discretization with K-medoids
No
Is discretization valid?
Yes
Formation of Bayesian Network structure
Node Probability Table
No
Is the model valid?
Yes
End

Figure 1. Research Methodology

3. Result and Discussion

3.1. Research Object

Data on damage to houses in Padang City due to the earthquake that hit West Sumatra on September 30, 2009, was obtained from the Regional Disaster Management Authority (RDMA) of Padang, and historical earthquake data from the Meteorological, Climatological, and Geophysical Agency (MCGA). Twenty-five thousand individual house-building data were used to create Bayesian networks. Five exogenous factors and three endogenous variables were employed in this investigation. Exogenous variables are not influenced by other variables, while other variables influence endogenous variables.

Table 1. Types of Research Variables

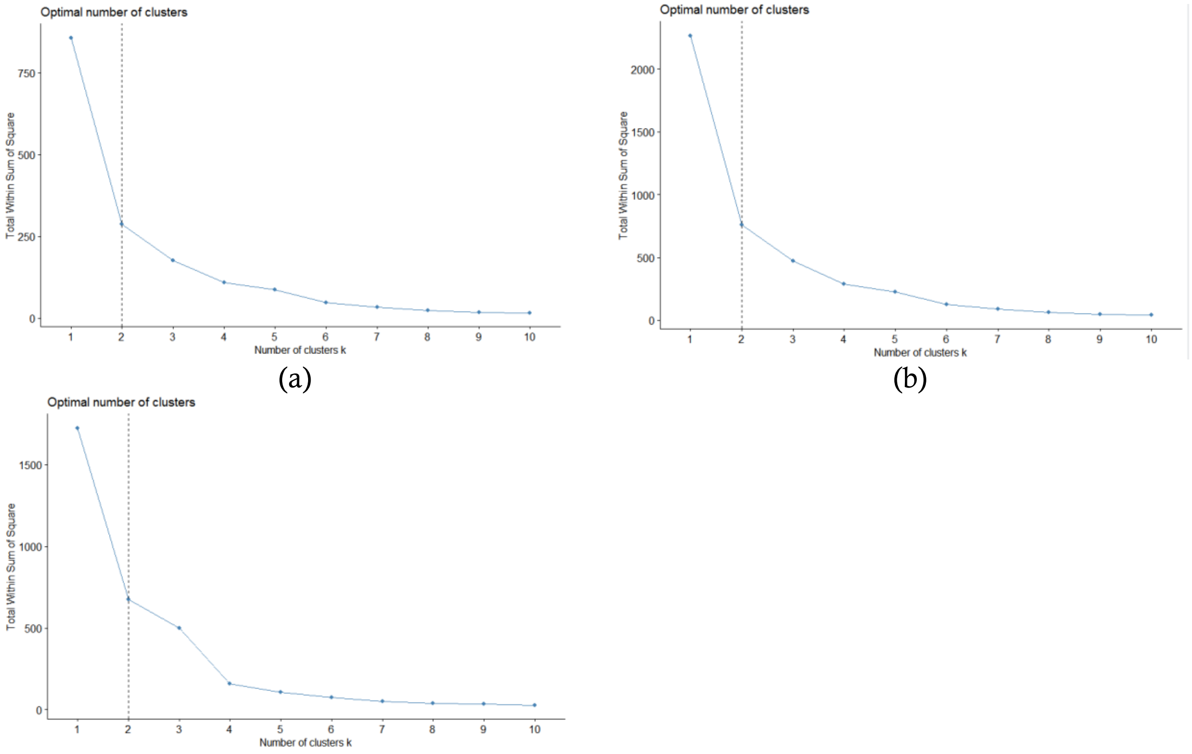
	Variable	**Type of Variable**
Exogenous Variable	Construction type (X_1)	Discrete
	Epicentral distance (X_3)	Continuous
	Soil type (X_4)	Discrete
	Slope (X_6)	Discrete
	Distance to fault (X_7)	Continuous
Endogenous Variable	Peak Ground Acceleration (X_2)	Continuous
	Landslide risk (X_5)	Discrete
	Damage rate (X_8)	Discrete

The variables were chosen based on past literature or study, including research undertaken by Bayraktarli et al. [13], [14] and Li et al. [15], [16]. Table 1 shows that there are three continuous variables in the research data there are Peak Ground Acceleration (X_2), epicentral distance (X_3), and distance to fault (X_7). The K-Medoids algorithm is used to discretize variables before using the Bayesian network to classify the level of damage to buildings.

3.2. K-Medoids Algorithm for Variable Discretization

In addition to being used for object clustering, the K-Medoids algorithm can also be used to discretize continuous data. The number of features in the discretization process is limited to only two. The first feature is a discretizable variable, and the second feature is an assumed constant auxiliary feature. If there are M objects in a set of objects $\boldsymbol{F} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m, \ldots, \boldsymbol{f}_M\}$ then the set of variables is $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2\}$,, where $\boldsymbol{X}_1 = \{f_{11}, \ldots, f_{21}, \ldots, f_{M1}\}$ and $\boldsymbol{X}_2$are constant. The following stage in the clustering procedure is the same as the phases in the K-Medoids method. The same category applies to objects in the same cluster.

However, before discretizing the variables, the best cluster determination for each variable must be found. The elbow approach was utilized to determine the ideal number of clusters in this investigation. For each K, the sum of square error (SSE) value is determined using the elbow method. The SSE had seen a significant decline and has the highest number of clusters [17]. The optimal number of clusters for each variable using the Elbow method, there are two clusters, could be deduced from Figure 2. After determining the best number of clusters for each variable, the variables are discretized using the K-Medoids algorithm. Figure 3 shows the discretization findings for each variable.

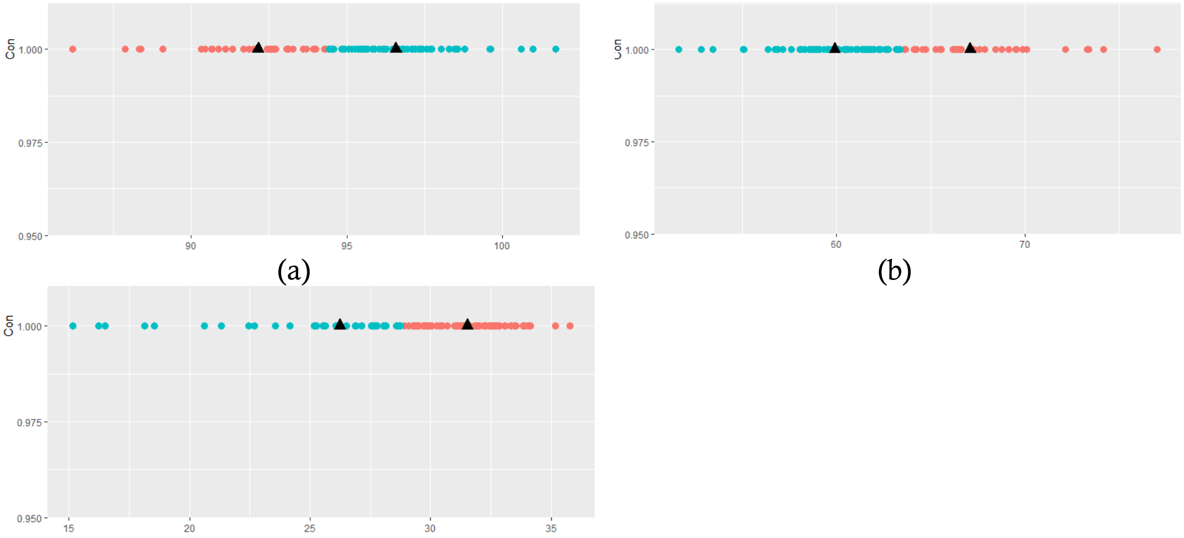


(a)

(b)

(c)

Figure 2. Determination of Optimal K using the Elbow Method for (a) Peak Ground Acceleration (b) Epicentral Distance (c) Distance to Fault



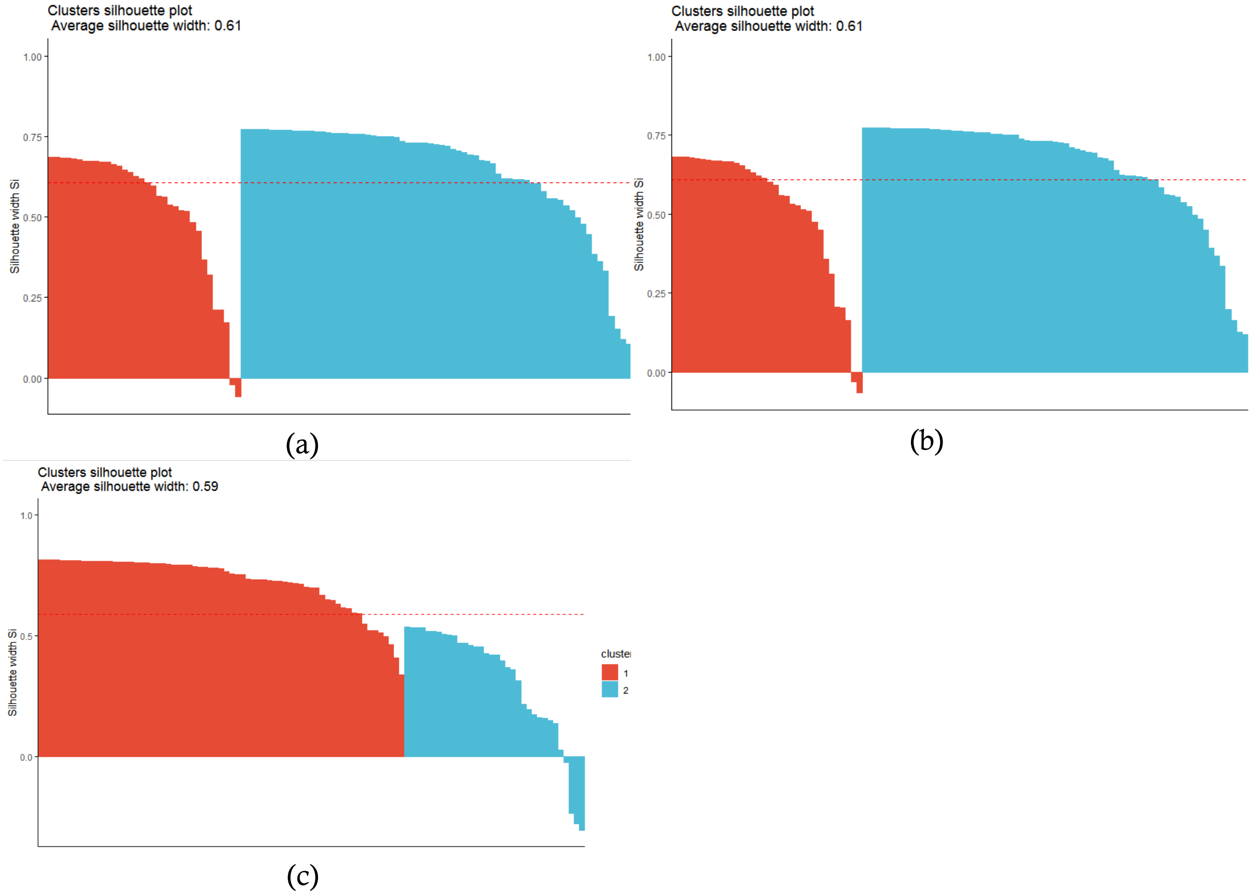
(a)

(b)

(c)

Figure 3. Discretized K-Medoids Results for (a) Peak Ground Acceleration (b) Epicentral Distance (c) Distance to Fault

After completing the grouping procedure as indicated in Figure 3, clusterization validation is performed, which is referred to as discretization validation in this case. The number of objects with a positive silhouette coefficient value is compared to the total number of objects for each variable at the variable discretization validation level (Figure 4). The discretization validation level in this scenario is 98 percent for the variable Peak Ground Acceleration (X_2) and epicentral distance (X_3), and 96 percent for the variable distance to fault (X_7).



(a)

(b)

(c)

Figure 4. Silhouette Coefficient Value for (a) Peak Ground Acceleration (b) Epicentral Distance (c) Distance to Fault

3.3. Formation of Bayesian Network Structure

Forming a BN structure is the next stage. Expert advice regarding numerous earlier scientific writings, including Bayraktarli et al. [13], [14] and Li et al. [15], [16], was used to form the BN structure in this study. Two key elements, exposure factors and system resilience factors impact the amount of damage to buildings caused by earthquakes [12]. Exposure considerations include magnitude, depth, epicentral distance, hypocentral distance, and other earthquake-related characteristics. The system resilience factor is linked to the environmental factors that cause disasters and the building's attributes. In 2012, Li focused his research on the extent of damage caused by earthquakes from a human perspective, which is influenced by exposure variables and system resilience factors.

Meanwhile, the research conducted by Bayraktarli [13], [14] only pays attention to the level of damage from the exposure factor. The relationship between variables and the probability for each variable can be seen in Figure 5. From each probability value of the level of damage, it can be concluded that the West Sumatra earthquake in 2009 caused damage to houses in the city of Padang, mostly at level two, which is moderate damage.

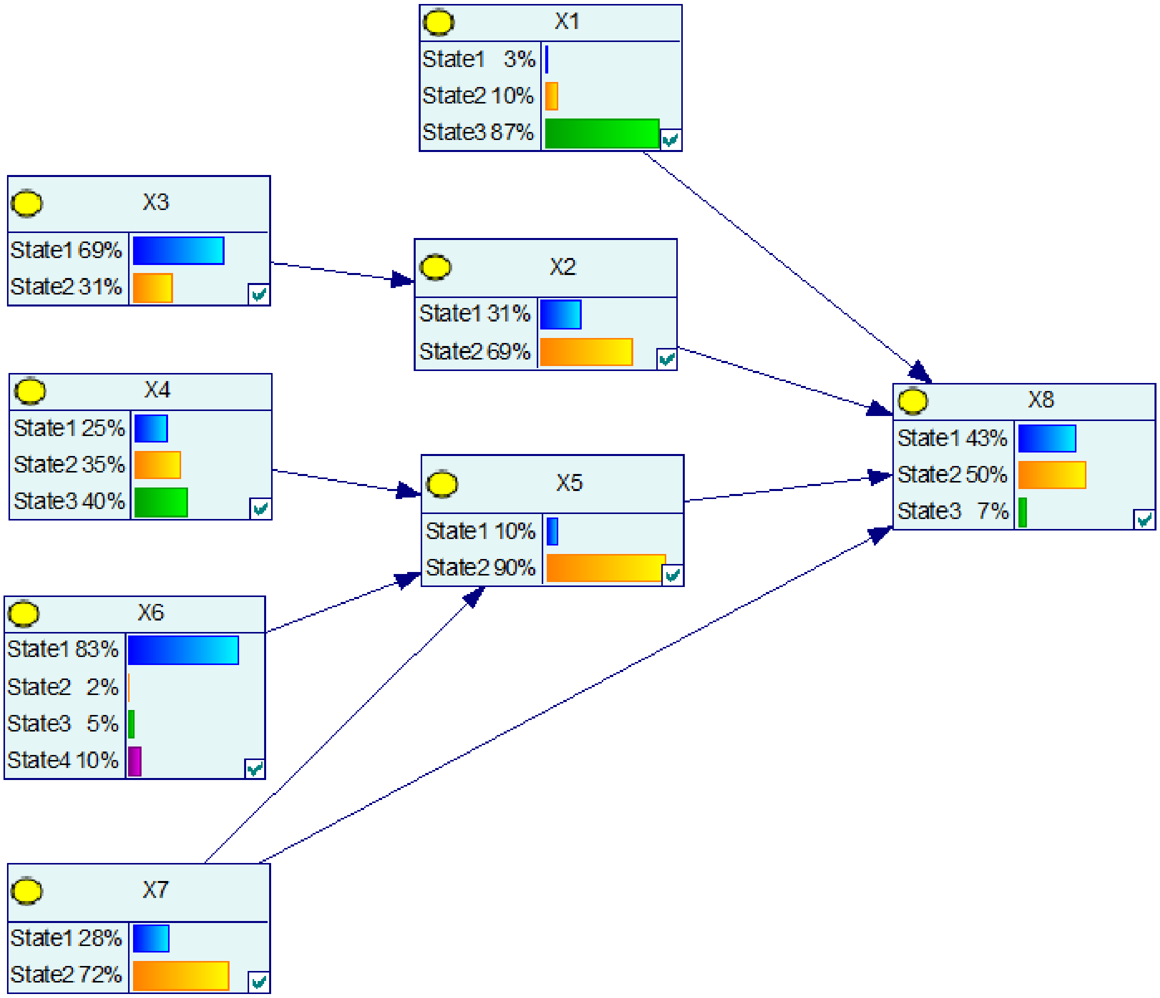


Figure 5. Structure of BN and NPT for Each Node

Table 2. Types of Research Variables

Damage Rate (X_8)	Actual			
Prediction	Slight (1)	Medium (2)	Heavy (3)	
Slight (1)	9457	552	482	10491
Medium(2)	16	13650	2	13668
Heavy(3)	34	121	686	841
	9507	14323	1170	25000

Furthermore, an assessment of the model's performance is carried out. The initial stage of performance evaluation is compiling a confusion matrix for the level of damage. The variable level of damage consists of three states, namely mild (state 1), moderate (state 2), and severe (state 3). In the confusion matrix, a comparison of the predicted and actual results is carried out, and complete details can be seen in Table 2.

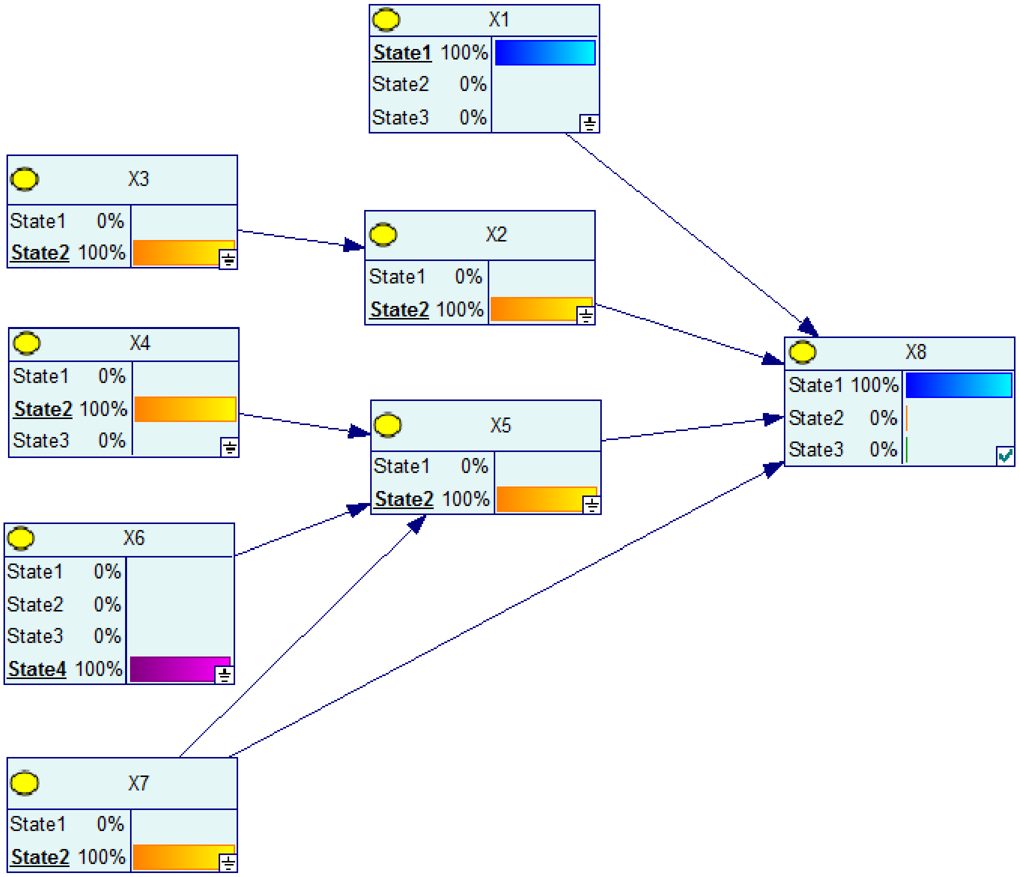


Figure 6. Structure of BN and NPT by Entering State of Variables

Then from the confusion matrix, the accuracy of each model is calculated by comparing the number of correct values with the amount of data. The level of model accuracy when using the K-Medoids discretization method is as follows,

$$\text{Accuracy rate } = \frac{9457 + 13650 + 686}{25000} = 95,17\%$$

The calculation results show that the model's accuracy rate is 95.17%.

Table 3. Position of Variable Determinants of Damage Level

Variable	State
Construction type (X_1)	1
Peak Ground Acceleration (X_2)	2
Epicentral distance (X_3)	2
Soil type (X_4)	2
Landslide risk (X_5)	2
Slope (X_6)	4
Distance to fault (X_7)	2

The BN model provides a probabilistic approach to obtaining an inference. Inference in a BN is obtained from the relationship of each node in the structure. Suppose the values of all the variables determining the level of damage are known, as shown in Table 3. Then, using the GeNIe software, the Bayesian network structure and the probability table for each variable are obtained, as shown in Figure 6. So, with the value of variables that affect the level of damage as shown in Table 3, it is most likely that the level of damage is in the low category with a probability of 100%.

4. Conclusion

Using the K-Medoids Algorithm, we create a BN model in terms of variable discretization in this study. The algorithm is changed by assuming that one of the features is variable and the rest are constant. We offer the BN modeling, which involves various factors, including construction type (X_1), Peak Ground Acceleration (X_2), epicentral distance (X_3), soil type (X_4), landslide risk (X_5), slope (X_6), and the distance to fault (X_7), in assessing the extent of building damage caused by earthquakes. We built the BN model as one of the earthquake catastrophe mitigation attempts based on current data. This model can assist the government or the community in reducing the danger of building damage. BN may be enlarged to satisfy all requirements with modest tweaks (adding more nodes and linkages and changes to the marginal and conditional probability tables). Like other decision systems, Bayesian networks are a helpful tool for calculating event probabilities because they are an ideal representation of prior causal knowledge and observed data.

5. Acknowledgments

The authors like to express gratitude to LPPM Universitas Negeri Padang for providing total funding for this study through Research Grants.

References

[1] J. Han, M. Kamber, and J. Pei. (2011). Data Mining: Concepts and Techniques, 3rd ed. *Morgan Kaufmann Publishers.*

[2] D. Berrar. (2018). Bayes theorem and naive bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, pp. 403–412,

[3] S. Kabir and Y. Papadopoulos. (2019). Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review, *Safety Science*, vol. 115, pp. 154–175.

[4] D. P. Sari, D. Rosadi, A. R. Effendie, and Danardono. (2018). Application of Bayesian network model in determining the risk of building damage caused by earthquakes, in *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, 2018, vol. 2018-Janua.

[5] D. P. Sari, D. Rosadi, A. R. Effendie, and D. Danardono. (2021). Discretization methods for bayesian networks in the case of the earthquake, *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 299–307.

[6] D. P. Sari, D. Rosadi, A. R. Effendie, and Danardono. (2019). Disaster mitigation solutions with Bayesian network, *AIP Conference Proceedings*, vol. 2192.

[7] D. P. Sari, D. Rosadi, A. R. Effendie, and Danardono. (2019). K-means and bayesian networks to determine building damage levels, *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 2, pp. 719–727.

[8] A. L. Madsen, C. J. Butz, J. S. Oliveira, and A. E. dos Santos. (2022). Simple Propagation with Arc-Reversal in Bayesian Networks, *Proceedings of Machine Learning Research*, vol. 72. PMLR, pp. 260–271.

[9] S. García, J. Luengo, J. A. Sáez, V. López, and F. Herrera. (2013).A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 734–750.

[10] X. S. Yang. (2019). Introduction to algorithms for data mining and machine learning, *Introduction to Algorithms for Data Mining and Machine Learning*, pp. 1–173.

[11] P. Arora, Deepali, and S. Varshney. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data, *Physics Procedia*, vol. 78, pp. 507–512.

[12] R. Nagarajan, M. Scutari, and S. Lèbre. (2013). Bayesian Networks in R: with Applications in Systems Biology. *Springer*. New York.

[13] Y. Y. Bayraktarli and M. H. Faber. (2011). Bayesian probabilistic network approach for managing earthquake risks of cities, *Georisk*, vol. 5, no. 1, pp. 2–24.

[14] Y. Y. Bayraktarli, U. Yazgan, A. Dazio, and M. H. Faber. (2006). Capabilities of the Bayesian probabilistic networks approach for earthquake risk management. vol. 120, pp. 3320–3344.

[15] L. F. Li, J. F. Wang, H. Leung, and S. Zhao. (2012). A Bayesian Method to Mine Spatial Data Sets to Evaluate the Vulnerability of Human Beings to Catastrophic Risk, *Risk Analysis*, vol. 32, no. 6, pp. 1072–1092.

[16] L. F. Li, J. F. Wang, and H. Leung. (2010). Using spatial analysis and Bayesian network to model the vulnerability and make insurance pricing of catastrophic risk, *International Journal of Geographical Information Science*, vol. 24, no. 12, pp. 1759–1784.

[17] K. Kaur and R. K. Dhaliwal, Dalvinder Singh; Vohra. (2019). Statistically Refining the Initial Points for K-Means Clustering Algorithm, *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 2, no. 11, pp. 2972–2977.