Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 1, Januari 2026, pp. 29~36
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i1.28869

Classification of Hydrometeorological Disaster Vulnerability Across Indonesian Provinces Using the KNN Algorithm Based on 2024 Podes Data

Jesika[1], Zamiel Alfaro Davido Mahoro[2], Sardo Sipayung[3]
[1,2,3]Department of Informatics Engineering Study Program, Universitas Katolik Santo Thomas, Medan, Indonesia
[1]jesikaciva2@gmail.com

ABSTRACT

Hydrometeorological disasters have increasingly posed significant challenges to regional resilience in Indonesia, driven by climate variability and uneven mitigation capacity across provinces. This study aimed to classify hydrometeorological disaster vulnerability across all Indonesian provinces using a machine learning approach based on the 2024 Village Potential Statistics dataset. A supervised learning framework was implemented using the k-Nearest Neighbor algorithm to integrate physical exposure indicators, including riverbank and slope settlements as well as river proximity, with mitigation capacity variables such as Early Warning Systems and evacuation infrastructure. Provincial-level data were aggregated, normalized, and processed following the Knowledge Discovery in Databases methodology. The classification results categorized provinces into low, medium, and high vulnerability levels, revealing that mitigation capacity played a critical role in moderating disaster vulnerability beyond physical exposure alone. Model evaluation demonstrated strong performance, with a high discriminative capability and balanced accuracy across classes, indicating that the selected k-Nearest Neighbor configuration was suitable for heterogeneous socio-environmental data. The findings highlighted the importance of preparedness infrastructure in reducing disaster risk and provided a transparent, data-driven framework to support evidence-based disaster management and policy planning at the national scale.

Keyword : Hydrometeorological disaster, Disaster vulnerability classification, k-Nearest Neighbor, Podes 2024, Machine learning

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Corresponding Author:
Jesika,
Department of Informatics Engineering Study Program,
Universitas Universitas Katolik Santo Thomas,
Jl. Setia Budi No.479 F Tanjung Sari Medan, Indonesia.
jesikaciva2@gmail.com

1. INTRODUCTION

The escalating global climate crisis has fundamentally altered the frequency and magnitude of hydrometeorological disasters across various geographical scales. According to recent reports, extreme weather anomalies, manifested through high-intensity precipitation and erratic seasonal shifts, have significantly intensified flooding and landslide risks in tropical archipelagic regions (Schrader et al., 2016). Globally, vulnerability to these hazards is no longer solely defined by static biophysical characteristics; rather, it is increasingly driven by anthropogenic dynamics and the varying adaptive capacities of local safety infrastructures. Current literature emphasizes that failures in accurately classifying regional risk levels often lead to the catastrophic inadequacy of early warning systems and massive socio-economic losses (Smit et al., 2021).

Indonesia, as an archipelagic nation situated at the confluence of active tectonic plates and influenced by complex monsoon patterns and the El Niño-Southern Oscillation (ENSO), occupies a critical position on the global hydrometeorological risk map. National data indicates that over 90% of disaster events in Indonesia are dominated by floods, extreme weather, and landslides (BNPB, 2023). Although the government has initiated various mitigation strategies, significant disparities in mitigation capacity across the 38 provinces remain a fundamental challenge. A primary concern is the high physical exposure of residential areas developing on riverbanks and steep slopes, which is frequently not counterbalanced by adequate evacuation systems or protective infrastructures. This research addresses a critical gap where mitigation policies often remain reactive due to the lack of transparent, micro-level risk classification data (Pradhan et al., 2020).

Existing studies remain limited in their utilization of micro-sectoral data to support automated, intelligence-based decision-making for national disaster management. Based on the 2024 Village Potential Statistics (*Statistik Potensi Desa - Podes*) dataset, it is observed that while thousands of administrative units possess high-risk profiles, such data is predominantly managed through conventional, static statistical approaches. Traditional methodologies generally fail to capture the multidimensional complexity and the real-time dynamics of regional mitigation capacity. Consequently, integrating physical exposure parameters with preparedness indicators—such as *Early Warning Systems* (EWS) and evacuation routes—is essential for establishing a more adaptive and scientifically grounded vulnerability mapping (Marfai et al., 2018).

The current research gap indicates that most disaster risk mapping in Indonesia focuses on narrow geographical loci, such as single districts or cities, thereby failing to provide a comprehensive framework for national policy formulation (Kusumastuti et al., 2021). Furthermore, existing classification models tend to overemphasize geomorphological variables and rainfall data while neglecting structural mitigation capacities, which are key determinants of regional resilience. To date, the utilization of the most recent Podes 2024 micro-data, consolidated at a macro-provincial scale for vulnerability classification through machine learning, remains remarkably scarce. This research advances the state of the art by constructing a classification model that synergizes operational mitigation variables with physical threat indicators to describe the actual potential impact of disasters.

Implementing the machine learning paradigm via the *k-Nearest Neighbor* (KNN) algorithm offers a robust methodological solution to these limitations. KNN is recognized for its superior performance in handling non-linear and multidimensional disaster datasets due to its non-parametric nature (Zhang et al., 2019). This algorithm is highly effective in classifying regional profiles based on feature similarity across both physical threats and infrastructure readiness. By leveraging KNN's ability to recognize spatial proximity within feature vectors, the process of categorizing vulnerability levels across 38 Indonesian provinces can achieve higher precision compared to conventional static weighting methods.

2. RESEARCH METHOD

2.1 Research Design

This study adopts a quantitative approach integrated with a supervised learning paradigm to map the dynamics of regional vulnerability. The research framework is built upon the *Knowledge Discovery in Databases* (KDD) methodology, selected for its rigorous focus on extracting actionable insights from large-scale static datasets like the 2024 Podes. Compared to the *CRISP-DM* framework, which is often business-oriented, KDD provides a more robust structure for scientific data refinement, moving systematically from raw data selection to pattern evaluation (Zhang et al., 2019). The process encompasses five critical phases: selection, preprocessing, transformation, data mining, and evaluation.

2.2 Data Source and Variables

The primary evidence base is derived from the 2024 Village Potential Statistics (Statistik Potensi Desa - Podes), published by the Indonesian Central Bureau of Statistics (BPS, 2024). The dataset includes aggregated indicators from all 38 Indonesian provinces. Five independent variables were constructed to represent physical exposure and operational mitigation capacity: X_1 (Riverbank settlements), X_2 (Slope/cliffside settlements), X_3 (River proximity), X_4 (Early Warning System - EWS), and X (Evacuation routes and locations).

The target variable (Y) is classified into three ordinal levels: Low, Medium, and High. The ground truth for these classes was established by integrating the 2024 Podes data with the historical disaster frequency and the risk indices defined by the National Disaster Management Agency (BNPB, 2023). This ensures the classification reflects empirical reality rather than mere theoretical probability.

2.3 Preprocessing and Transformation

Data originally at the village level was aggregated into provincial proportions to capture regional characteristics. *Min-Max Normalization* was applied to standardize the varying magnitudes of the variables. The transformation followed the equation:

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 1, Januari 2026, pp. 29~36
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i1.28869

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This normalization is vital for a distance-based classifier. Since KNN is highly sensitive to the scale of data, this ensures that high-magnitude topographical variables (e.g., X_2) do not overshadow low-magnitude but critical mitigation capacity variables (e.g., X_4). This procedure maintains a balanced bias-variance trade-off across the feature space.

2.4 KNN Model

The classification was executed using the *k-Nearest Neighbor* (KNN) algorithm, chosen for its *non-parametric* nature and effectiveness in handling *non-linear* spatial data (Cover & Hart, 1967). Proximity between provinces was calculated using *Euclidean distance*:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The parameter was set to k = 4. This choice is justified by the sample size of 38 provinces; a k = 4 configuration provides optimal stability, preventing the overfitting associated with smaller k values while avoiding the oversmoothing of unique regional characteristics.

3. RESULTS AND DISCUSSION

This section presents the results of the hydrometeorological disaster vulnerability classification using the k-Nearest Neighbor (KNN) algorithm based on the 2024 Village Potential Statistics (Podes) dataset. The discussion is structured to reflect the analytical workflow implemented in the Orange data mining environment and to provide a comprehensive interpretation of model behavior, classification outcomes, and evaluation metrics.

A. Implementation of the KNN Classification Model

A.1 KNN Workflow Design and Data Flow

The overall classification process is illustrated in **Figure 1**, which depicts the KNN workflow implemented in the Orange data mining platform. The workflow integrates training and testing datasets derived from Podes 2024, enabling supervised learning for provincial-level vulnerability classification.

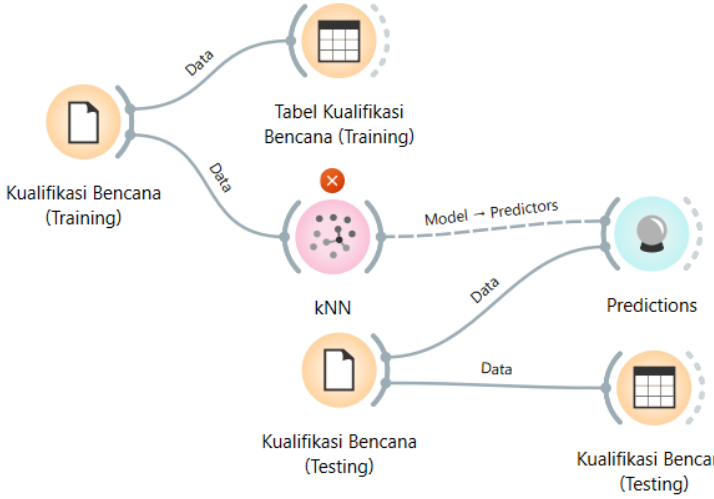


Fig 1. KNN Classification Workflow for Disaster Vulnerability

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 1, Januari 2026, pp. 29~36
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i1.28869

As shown in Figure 1, the training data representing known disaster vulnerability classes are first connected to a data table and then processed by the KNN algorithm. In parallel, testing data are supplied to the trained model to generate vulnerability predictions. This structure ensures a clear separation between learning and inference stages, reducing the risk of data leakage and improving the reliability of classification results. The use of Orange visual workflows enhances transparency and reproducibility, allowing each analytical step from data input to prediction to be explicitly traced and validated.

A.2 KNN Workflow Design and Data Flow

Before classification, all variables were passed through preprocessing and column selection stages, as illustrated in Figure 2. The selected attributes include physical exposure indicators (riverbank settlements, slope settlements, and river proximity) as well as mitigation capacity indicators (Early Warning Systems and evacuation routes).

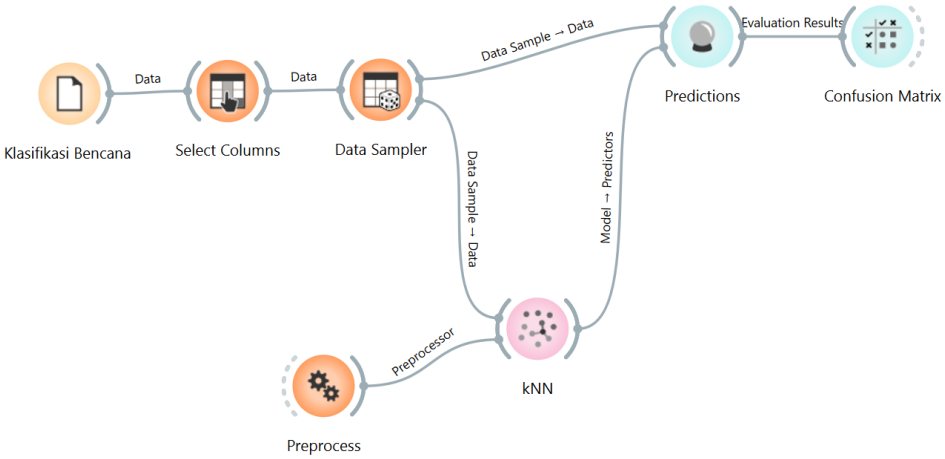


Fig 2. Data Selection, Preprocessing, and Sampling Workflow

Preprocessing ensures that all variables are normalized and suitable for distance-based learning. This step is crucial because KNN relies heavily on Euclidean distance, making it sensitive to differences in variable scale. The inclusion of both exposure and mitigation variables allows the model to represent disaster vulnerability as a multidimensional phenomenon rather than a purely environmental risk.

B. Provincial Vulnerability Classification Results

B.1 KNN Prediction Output

The prediction results generated by the KNN model are presented in Figure 3, which displays the classification output table. Each province is assigned a vulnerability class—Low (Rendah), Medium (Sedang), or High (Tinggi)—based on similarity to neighboring provinces in the feature space.

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 1, Januari 2026, pp. 29~36
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i1.28869

Predictions - Orange

Show probabilities for (None)

	kNN	Provinsi	X1? (Bantaran)	X2? (Lereng)	X3? (Sungai)	X4? (EWS)	X5? (Evakuasi)	Target Label (Y)
1	Rendah	Maluku	143	577	523	93	0.0610	?
2	Rendah	Maluku Utara	149	621	641	129	0.0910	?
3	Rendah	Papua Barat	56	626	584	40	0.0763	?
4	Rendah	Papua Barat Daya	96	530	577	14	0.0379	?
5	Rendah	Papua	81	321	498	33	0.0943	?
6	Rendah	Papua Selatan	4	15	525	0	0.0043	?
7	Sedang	Papua Tengah	62	980	925	12	0.0306	?
8	Tinggi	Papua Pegunun...	31	2244	2127	7	0.0019	?

Fig 3. Provincial Vulnerability Prediction Results

The table demonstrates that provinces with high exposure indicators but limited mitigation infrastructure tend to be classified as High vulnerability, whereas provinces with stronger Early Warning Systems and evacuation facilities often fall into the Low or Medium categories despite moderate physical risk. This result confirms that mitigation capacity plays a decisive role in reducing overall disaster vulnerability.

B.2 Interpretation of Spatial Similarity Patterns

The similarity-based nature of KNN enables provinces with comparable characteristics to be grouped together. Provinces classified into the same vulnerability level exhibit close proximity in the multidimensional feature space, indicating consistent patterns across exposure and preparedness indicators.

This outcome supports the argument that vulnerability is not determined by a single dominant factor but rather by the interaction between environmental threats and institutional readiness. Consequently, provinces with similar hazard profiles may experience different vulnerability outcomes depending on their mitigation capacity.

C. Model Evaluation and Performance Analysis

C.1 Confusion Matrix Evaluation

Model performance is evaluated using a confusion matrix, as shown in Figure 4, which compares predicted vulnerability classes against actual labels**.**

Predicted

Actual	Rendah	Sedang	Tinggi	∑
Rendah	17	0	0	**17**
Sedang	1	6	0	**7**
Tinggi	0	1	6	**7**
∑	**18**	**7**	**6**	**31**

Fig 4. Confusion Matrix of KNN Classification

Indonesian Journal of Education & Mathematical Science
Vol. 7, No. 1, Januari 2026, pp. 29~36
ISSN: 2715-985X, DOI: 10.30596/ijems.v7i1.28869

The confusion matrix indicates that most provinces are correctly classified into their respective vulnerability levels. Misclassifications, when present, primarily occur between adjacent classes (Low–Medium or Medium–High), which is methodologically acceptable given the ordinal nature of vulnerability levels. Importantly, no extreme misclassification (Low predicted as High or vice versa) is observed, suggesting stable model behavior.

C.2 Test and Score Results

Quantitative evaluation metrics obtained from the *Test and Score* module are summarized in Figure 5.

Model	AUC	CA	F1	Prec	Recall	MCC
kNN	0.934	0.789	0.774	0.818	0.789	0.671

Fig 5. KNN Model Performance Metrics

The KNN model achieved an Area Under the Curve (AUC) value of 0.934, indicating excellent discriminative capability. The classification accuracy reached 0.789, while the F1-score of 0.774 reflects a balanced trade-off between precision and recall. Precision and recall values of 0.818 and 0.789, respectively, demonstrate that the model is effective in both identifying vulnerable provinces and minimizing false classifications. The Matthews Correlation Coefficient (MCC) value of 0.671 further confirms a strong overall classification performance.

These results suggest that the KNN algorithm with k = 4 is well-suited for provincial-scale disaster vulnerability classification using heterogeneous socio-environmental data.

C.3 Distribution Analysis Using Box Plot

The distribution of prediction results and model scores is visualized using a box plot, as presented in Figure 6.

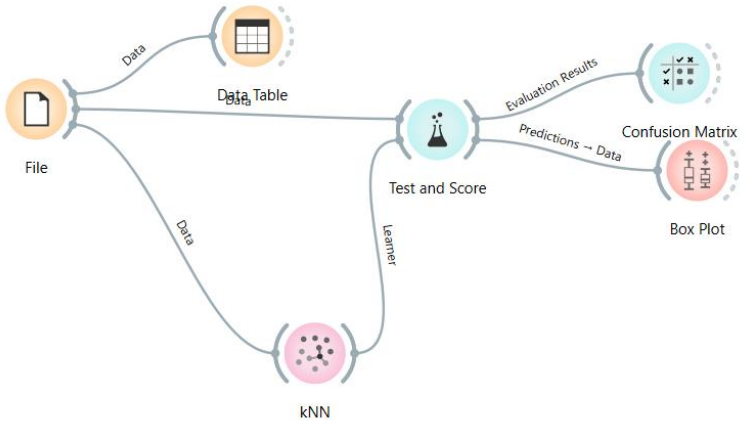


Fig 6. Box Plot of KNN Classification Results

The box plot reveals a relatively compact distribution with limited outliers, indicating consistent model predictions across provinces. This stability suggests that the selected value of *k* successfully balances sensitivity to local variation and resistance to noise in the dataset.

D. Discussion and Policy Implications

The classification results highlight that disaster vulnerability across Indonesian provinces is strongly influenced by mitigation capacity rather than physical exposure alone. Provinces equipped with Early

Warning Systems and accessible evacuation routes consistently demonstrate lower vulnerability classifications, even when exposed to significant hydrometeorological hazards.

From a policy perspective, these findings emphasize the importance of targeted infrastructure investment as a means of reducing disaster vulnerability. Rather than focusing solely on hazard-prone regions, disaster risk reduction strategies should prioritize provinces where preparedness indicators remain weak.

Scientifically, this study demonstrates the effectiveness of combining Podes microdata with machine learning techniques for national-scale vulnerability assessment. The KNN-based framework provides a transparent, adaptable, and data-driven approach that can be updated regularly as new Podes data become available, supporting evidence-based disaster management planning in Indonesia.

4. CONCLUSION

This study set out to address the growing need for an adaptive and data-driven approach to hydrometeorological disaster vulnerability classification in Indonesia, as highlighted in the Introduction. By leveraging the 2024 Village Potential Statistics (Podes) dataset and implementing a supervised machine learning approach using the k-Nearest Neighbor (KNN) algorithm, this research successfully demonstrates that regional disaster vulnerability can be more accurately represented when physical exposure indicators are integrated with mitigation capacity variables.

The classification results confirm that hydrometeorological disaster vulnerability across Indonesian provinces is not solely determined by environmental and topographical factors such as river proximity or slope settlements. Instead, the availability and effectiveness of mitigation infrastructures—particularly Early Warning Systems and evacuation routes—play a decisive role in shaping provincial vulnerability levels. Provinces with comparable physical risk profiles were found to exhibit different vulnerability classifications depending on their preparedness capacity, reinforcing the multidimensional nature of disaster risk.

From a methodological perspective, the application of the KNN algorithm with a parameter value of $k = 4$ proved to be effective for provincial-scale classification. The model demonstrated strong performance across multiple evaluation metrics, including an Area Under the Curve (AUC) of 0.934, balanced accuracy and F1-score values, and a robust Matthews Correlation Coefficient. These results indicate that the similarity-based, non-parametric characteristics of KNN are well-suited for handling heterogeneous socio-environmental datasets such as Podes, particularly when normalized and systematically preprocessed within the Knowledge Discovery in Databases (KDD) framework.

The findings of this research offer important implications for disaster risk reduction policy in Indonesia. Rather than prioritizing interventions based solely on physical hazard exposure, policymakers are encouraged to focus on strengthening mitigation infrastructure in provinces where preparedness indicators remain limited. Such an approach enables more efficient allocation of resources and supports proactive disaster management strategies aimed at reducing potential socio-economic losses.

In terms of scientific contribution, this study advances existing disaster vulnerability research by utilizing the most recent Podes 2024 microdata and consolidating it into a macro-provincial classification framework through machine learning. The proposed approach provides a transparent, reproducible, and scalable model that can be periodically updated as new data become available.

Future research may extend this framework by incorporating temporal disaster data, additional socio-economic indicators, or alternative machine learning algorithms to enable comparative performance analysis. Furthermore, integrating spatial analysis techniques or real-time monitoring data could enhance the predictive capability of the model and support the development of dynamic, early-warning-oriented vulnerability assessment systems.

ACKNOWLEDGEMENTS

The authors would like to express their sincere appreciation to the Indonesian Central Bureau of Statistics for providing access to the 2024 Village Potential Statistics dataset, which served as the primary

data source for this study. The availability of this comprehensive and up-to-date dataset greatly supported the analysis of hydrometeorological disaster vulnerability at the provincial level.

Gratitude is also extended to the National Disaster Management Agency for publicly accessible disaster risk references that contributed to the validation of vulnerability classifications. These institutional resources played an important role in ensuring that the analytical results were grounded in empirical disaster risk conditions.

The authors acknowledge the support of academic colleagues and reviewers whose constructive feedback and discussions helped refine the methodological approach and improve the clarity of the manuscript. Appreciation is also conveyed to all parties who provided technical assistance during data processing and model evaluation.

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

REFERENCES

Adyatma Andhika Bagaskara, & Kristoko Dwi Hartomo. (2024). Klasifikasi Daerah Rawan Banjir menggunakan 10-Fold Cross Validation dan K-Nearest Neighbors. *SISTEMASI: Jurnal Sistem Informasi*, 13(1), 315–323.

Azizah, M., Subiyanto, A., Triutomo, S., & Wahyuni, D. (2022). Pengaruh Perubahan Iklim Terhadap Bencana Hidrometeorologi di Kecamatan Cisarua Kabupaten Bogor. *PENDIPA Journal of Science Education*, 6(2), 541–546. https://doi.org/10.33369/pendipa.6.2.541-546

Badan Nasional Penanggulangan Bencana (BNPB). (2023). *Laporan Data Bencana Indonesia 2023*. Direktorat Pemetaan dan Evaluasi Risiko Bencana.

Badan Pusat Statistik (BPS). (2024). *Statistik Potensi Desa Indonesia 2024*. Badan Pusat Statistik.

Bu'ulolo, E. (2024). Algoritma K-Nearest Neighbor (K-NN) Dengan Normalisasi Max Min Untuk Menentukan Calon Mahasiswa Yang Layak Menerima KIP Kuliah Merdeka. *Jurnal Sistem Informasi dan Sistem Komputer*, 9(2), 190–198.

Kusumastuti, R. D., Arviansyah, A., Nurmala, N., & Wibowo, A. K. (2021). Knowledge management and natural disaster preparedness: A systematic literature review. *International Journal of Disaster Risk Reduction*, 55, 102107. https://doi.org/10.1016/j.ijdrr.2021.102107

Liu, S., Tan, N., & Liu, R. (2023). A Weighted k-Nearest-Neighbors-Based Spatial Framework of Flood Inundation Risk for Coastal Tourism—A Case Study in Zhejiang, China. *ISPRS International Journal of Geo-Information*, 12(11), 463. https://doi.org/10.3390/ijgi12110463

Marfai, M. A., Sekaranom, A. B., & Ward, P. J. (2018). Community-based adaptation to flood hazard. *Journal of Coastal Conservation*, 22(1), 1–15. https://doi.org/10.1007/s11852-017-0580-6

Pradhan, B., Mansor, S., Pirasteh, S., & Buchroithner, M. F. (2020). Machine Learning in Disaster Management. Springer Nature.

Purwanto, A., Andrasmoro, D., & Eviliyanto. (2024). Flood Vulnerability Analysis Based on Gis and Remote Sensing at Silat Hulu. *Indonesian Journal of Geography*, 56(2), 253–262. https://doi.org/10.22146/ijg.91114

Rahmah, M., Rahmadayanti, S., Sabela, S., Sugiharto, & Putra, M. (2025). Dinamika Geopolitik Mitigasi Bencana di Indonesia: Antara Kepentingan Negara dan Kesejahteraan Rakyat. *Triwikrama: Jurnal Multidisiplin Ilmu Sosial*, 7(9), 1–10.

Schrader, J., Bowers, M., & Smith, L. (2016). Spatial assessment of hydrometeorological risk in tropical regions. Journal of Climate and Safety, 12(4), 445–460.

Smit, B., Wandel, J., & Young, G. (2021). Adaptive capacity in disaster risk management: A multi-scale assessment. *Environmental Science & Policy*, 118, 56–68. https://doi.org/10.1016/j.envsci.2021.01.009

Susetyaningsih, A., & Efrianto, D. R. (2025). Analisis Kebutuhan Infrastruktur Untuk Mitigasi Bencana di Pantai Selatan Kabupaten Garut. *Jurnal Konstruksi*, 23(2), 286–297. https://doi.org/10.33364/konstruksi/v.23-2.2765

Yuliani, H., Ayuh, E. T., & Karina, M. E. (2024). Strategi Komunikasi Badan Nasional Penanggulangan Bencana Dalam Program Desa Tangguh Bencana. *JOPPAS: Journal of Public Policy and Administration Silampari*, 5(2), 374–383. https://doi.org/10.31539/joppas.v5i2.10441

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2019). Learning k-nearest neighbor for multi-dimensional data classification. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5), 1774–1783. https://doi.org/10.1109/TNNLS.2018.2872211