

Pemodelan Prediksi Magnitudo Gempa Bumi di Indonesia Periode 2023–2025 Menggunakan Algoritma Naïve Bayes

M. Faris Daffarindra^{1*}, Akhiyar Waladi², Pradita Eko Prasetyo Utomo³, Ulfa Khaira⁴

^{1,2,3,4}Universitas Jambi

*Penulis Korespondensi : faris.6dsdiaf@gmail.com

ABSTRACT

Earthquake prediction is critical for disaster mitigation in Indonesia, one of the world's most seismically active countries. This study applies the Gaussian Naïve Bayes algorithm to classify earthquake magnitudes using BMKG data for 2023–2025. After preprocessing 30,000 records, 23,979 valid entries remained with six predictive variables: latitude, longitude, depth, phasecount, azimuth gap, and magnitude type. Magnitudes were categorized into Light (< 3.0), Moderate ($3.0-4.9$), and Strong (≥ 5.0). The model achieved 71.68% overall accuracy, with precision of 0.77 and recall of 0.66 for the Moderate class, and recall of 0.78 for the Light class. Five-fold cross-validation yielded a mean accuracy of 68.78% ($\pm 7.32\%$). Comparative evaluation showed Random Forest (83.63%) and SVM (79.15%) outperformed Naïve Bayes overall; however, only Naïve Bayes detected the Strong class (precision = 0.18; recall = 0.25). The novelty lies in the three-algorithm comparative evaluation and critical assessment of the independence assumption against real seismic data. The model is suitable as a preliminary screening tool for early warning systems.

Article History

Received : 16-12-2025
Revised : 04-07-2026
Accepted : 06-07-2026

Keywords

Gempa Bumi
Klasifikasi
Naïve Bayes
Prediksi

ABSTRAK

Prediksi gempa bumi sangat penting untuk mitigasi bencana di Indonesia sebagai salah satu negara paling aktif seismik di dunia. Penelitian ini menerapkan algoritma Gaussian Naïve Bayes untuk mengklasifikasikan magnitudo gempa menggunakan data BMKG periode 2023-2025. Setelah pra-pemrosesan 30.000 rekaman, diperoleh 23.979 data valid dengan enam variabel prediktor: lintang, bujur, kedalaman, phasecount, azimuth gap, dan jenis magnitudo. Magnitudo dikelompokkan menjadi Ringan ($< 3,0$), Sedang ($3,0-4,9$), dan Kuat ($\geq 5,0$). Model mencapai akurasi 71,68% dengan precision 0,77 dan recall 0,66 pada kelas Sedang, serta recall 0,78 pada kelas Ringan. Validasi 5-fold cross-validation menghasilkan rata-rata akurasi 68,78% ($\pm 7,32\%$). Perbandingan dengan Random Forest (83,63%) dan SVM (79,15%) menunjukkan hanya Naïve Bayes yang mampu mendeteksi kelas Kuat (precision = 0,18; recall = 0,25). Kebaruan penelitian ini terletak pada evaluasi komparatif tiga algoritma dan analisis kritis asumsi independensi fitur Naïve Bayes terhadap data seismik nyata. Model ini layak sebagai alat skrining awal dalam sistem peringatan dini.

PENDAHULUAN

Indonesia berada di pertemuan beberapa lempeng tektonik besar, yaitu Lempeng Eurasia, Pasifik, dan Indo-Australia, sehingga memiliki tingkat kerawanan seismik yang sangat tinggi. Frekuensi gempa bumi yang tinggi mendorong perlunya penelitian mendalam mengenai karakteristik gempa, perkiraan magnitudo, serta strategi mitigasi bencana. Penelitian tentang aktivitas seismik di berbagai wilayah Indonesia menunjukkan bahwa kompleksitas struktur lempeng menghasilkan karakteristik gempa yang beragam dengan dampak signifikan pada wilayah padat penduduk [1].

Sejumlah penelitian terdahulu telah mengeksplorasi pendekatan *machine learning* untuk prediksi dan klasifikasi gempa bumi berbasis data. Jena et al. [2] mengintegrasikan *neural network* dan Analytic Hierarchy Process (AHP) untuk penilaian risiko seismik perkotaan, sementara Asim et al. [3] menggunakan adaptive neuro-fuzzy inference system untuk prediksi probabilitas gempa di kawasan Hindukush. Di dalam negeri, Sari dan Hariadi [4] menerapkan *Decision Tree* untuk klasifikasi magnitudo gempa di Pulau Jawa. Maulita dan Wahid [5] membandingkan lima algoritma

ML, yaitu Random Forest, SVR, XGBoost, LightGBM, dan MLP, untuk prediksi magnitudo gempa di Indonesia menggunakan fitur kedalaman dan geolokasi, dengan LightGBM memberikan performa terbaik (MAE = 0,47). Penelitian pada domain lain juga menunjukkan konsistensi Naïve Bayes sebagai pengklasifikasi yang efisien; Ariadi [6] membandingkan C4.5 dan Naïve Bayes dalam prediksi kredit dan memperoleh akurasi Naïve Bayes 89,52%, sementara Kusuma et al. [7] menerapkan Naïve Bayes untuk klasifikasi penyakit endokrin dengan akurasi ~84%. Meskipun demikian, penelitian yang secara khusus menerapkan Gaussian Naïve Bayes pada data BMKG skala besar untuk periode 2023-2025 masih terbatas, terutama yang mengevaluasi asumsi independensi fitur secara kritis dan membandingkan kemampuan deteksi kelas magnitudo ekstrem secara komparatif [8][9].

Berdasarkan latar belakang tersebut, rumusan masalah penelitian ini adalah: (1) Bagaimana performa algoritma Gaussian Naïve Bayes dalam mengklasifikasikan magnitudo gempa bumi berdasarkan data BMKG 2023-2025? (2) Seberapa besar pengaruh ketidakseimbangan kelas terhadap akurasi prediksi? (3) Apakah asumsi independensi fitur Naïve Bayes sesuai dengan karakteristik data seismik yang saling berkorelasi? Penelitian ini bertujuan: (1) membangun model klasifikasi Gaussian Naïve Bayes pada dataset BMKG 2023-2025, (2) mengevaluasi performa model menggunakan precision, recall, dan F1-score, (3) membandingkan performa dengan Random Forest dan SVM, serta (4) memberikan rekomendasi teknis peningkatan model untuk penelitian lanjutan.

METODE

Sumber Data

Penelitian ini menggunakan pendekatan kuantitatif prediktif dengan data yang bersumber dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) yang dipublikasikan melalui platform Kaggle oleh Kekavigi [10]. Dataset awal memiliki 131.830 baris dan 37 kolom. Atribut utama yang digunakan mencakup waktu kejadian (*datetime*), koordinat (*latitude*, *longitude*), magnitudo, jenis magnitudo (*mag_type*), kedalaman gempa (*depth*), jumlah fase gelombang seismik (*phasecount*), dan celah azimuth (*azimuth_gap*). Data diurutkan berdasarkan *datetime* secara menurun dan dipilih 30.000 baris yang merepresentasikan periode 1 Januari 2023 hingga 20 September 2025.

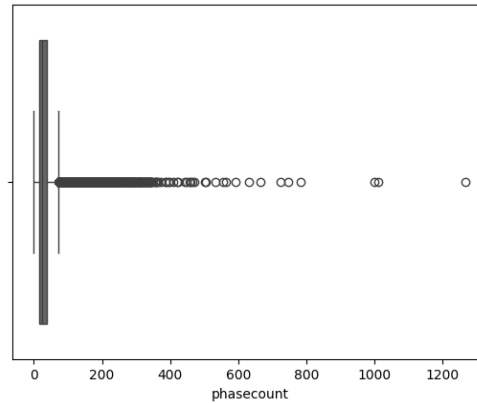
eventID	datetime	latitude	longitude	magnitude	mag_type	depth	phasecount	azimuth_gap	location	...	doublecouple	clvd	strikeNP1	dipNP1	rakeNP1	strikeNP2	dipNP2	rakeNP2	azgapM	misfit
0	bmg2008vkye 2008-11-01 00:31:25.143741+00:00	-0.04440	98.895531	2.989742	MLv	20	6.0	146.70738	Southern Sumatra, Indonesia	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	bmg2008vlag 2008-11-01 01:34:29.660856+00:00	-6.611860	129.387220	5.507549	mb	30	62.0	45.46764	Banda Sea	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	bmg2008vlaj 2008-11-01 01:38:14.802129+00:00	-3.650586	127.990680	3.539674	MLv	5	4.0	331.97404	Seram, Indonesia	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	bmg2008vibt 2008-11-01 02:20:05.909516+00:00	-4.198925	128.097000	2.424314	MLv	5	5.0	326.37993	Banda Sea	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	bmg2008vlcd 2008-11-01 02:32:18.756165+00:00	-4.091891	128.200470	2.410045	MLv	10	5.0	314.65474	Banda Sea	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Gambar 1. Tampilan Dataset Awal menggunakan fungsi `df.head()`

Tahapan Pra-Pemrosesan Data

Pra-pemrosesan mencakup tiga tahap. Pertama yaitu penanganan missing values, variabel numerik diisi menggunakan median dan variabel kategorikal menggunakan modus. Kedua yaitu penghapusan outlier menggunakan metode Interquartile Range (IQR) pada kolom *phasecount* (batas bawah = -18,50; batas atas = 73,50) dan *depth* (batas bawah = -38,00; batas atas = 90,00), sehingga tersisa 23.979 data valid. Ketiga yaitu transformasi fitur, *mag_type* dikodekan menggunakan *Label Encoding*, sedangkan fitur numerik dinormalisasi menggunakan *Min-Max Scaler* ke rentang 0-1.

Data dibagi dengan proporsi 80:20 menggunakan *stratified split*, menghasilkan 19.183 data latih dan 4.796 data uji.



Gambar 2. Boxplot distribusi *phasecount* sebelum penghapusan *outlier*

Nilai *magnitude* digunakan sebagai variabel target yang diklasifikasikan menjadi tiga kategori:

- Ringan : $magnitude < 3.0$
- Sedang : $3.0 \leq magnitude < 5.0$
- Kuat : $magnitude \geq 5.0$

Pengelompokan tersebut mempermudah interpretasi hasil model dan memungkinkan analisis distribusi gempa antar kategori secara statistik.

Penerapan Naïve Bayes

Proses pemodelan dalam penelitian ini menggunakan algoritma *Gaussian Naïve Bayes*, yaitu metode probabilistik berbasis Teorema Bayes yang mengasumsikan independensi antar variabel prediktor. Metode *Naïve Bayes* adalah teknik klasifikasi berbasis pendekatan statistik yang berfungsi untuk menentukan probabilitas suatu data termasuk ke dalam kelas tertentu. Model ini dipilih karena kemampuannya dalam menangani dataset berukuran besar dengan komputasi yang efisien, efektivitasnya pada klasifikasi multi-kelas [7][11]. Model dibangun menggunakan fungsi *GaussianNB()* dari *library scikit-learn* pada Python.

Persamaan umum yang digunakan dalam model adalah:

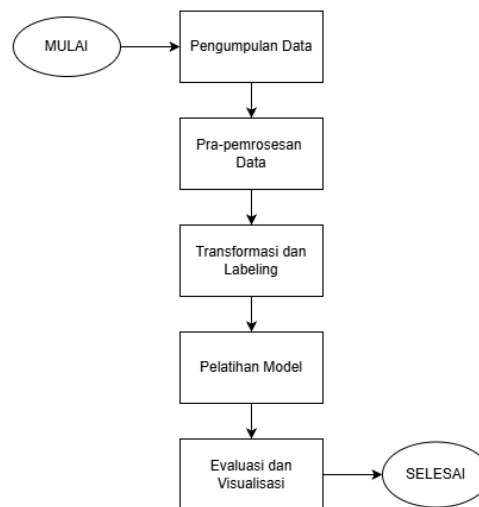
$$P(C_i | X) = \frac{P(X | C_i) \cdot P(C_i)}{P(X)}$$

Dengan,

- $P(C_i | X)$ = probabilitas kelas C_i terhadap atribut X ,
- $P(X | C_i)$ = probabilitas kemunculan atribut X pada kelas C_i ,
- $P(C_i)$ = probabilitas awal (prior) kelas C_i ,
- $P(X)$ = probabilitas total atribut X .

Satu catatan penting: asumsi independensi antar fitur tidak sepenuhnya berlaku pada data seismik, sebab *depth*, *azimuth_gap*, dan *phasecount* saling berkorelasi. Namun Hand & Yu [11] sudah menunjukkan bahwa Naïve Bayes tetap bisa efektif memprediksi label kelas meski asumsi ini dilanggar. Kestabilan model diuji lewat 5-fold cross-validation. Sebagai pembandingan, dipakai Random Forest ($n_estimators = 100$) dan SVM (*kernel RBF*), keduanya dipilih karena relevansinya dalam riset prediksi gempa sebelumnya [3][5]. Seluruh model dievaluasi dengan metrik akurasi, precision, recall, dan F1-score.

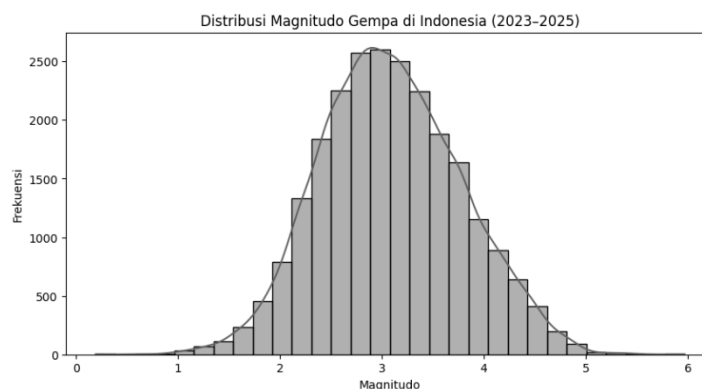
Alur keseluruhan proses penelitian digambarkan pada Gambar 3, yang terdiri dari lima tahap utama, yaitu pengumpulan data, pra-pemrosesan, transformasi dan labeling, pelatihan model, dan evaluasi hasil.



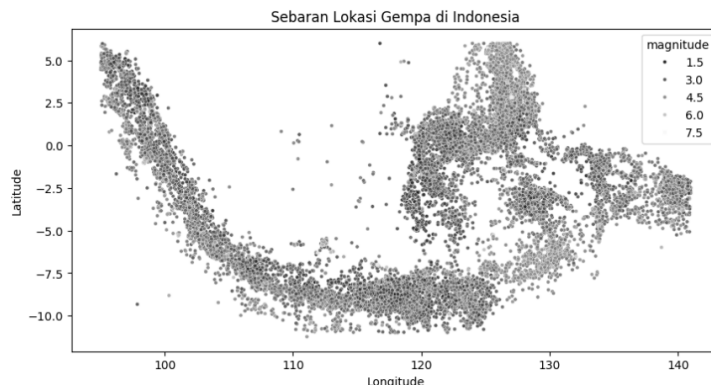
Gambar 3. Diagram alur penelitian prediksi magnitudo gempa bumi

Visualisasi Data

Visualisasi berperan penting dalam memahami pola distribusi gempa bumi dan kinerja model klasifikasi. Penelitian ini menghasilkan dua visualisasi utama:



Gambar 4. Histogram distribusi magnitudo gempa bumi di Indonesia tahun 2023-2025.



Gambar 5. Peta sebaran lokasi gempa bumi berdasarkan koordinat lintang dan bujur.

HASIL DAN PEMBAHASAN

Evaluasi Model Gaussian Naïve Bayes

Setelah melalui tahap pra-pemrosesan, diperoleh data dari yang awalnya 30.000 data, kini hanya menjadi 23.979 data. Model dibangun dengan variabel prediktor berupa *latitude*, *longitude*, *depth*, *phasecount*, *azimuth_gap*, serta *mag_type_encoded*, sedangkan variabel target adalah kategori magnitudo (*Ringan*, *Sedang*, *Kuat*).

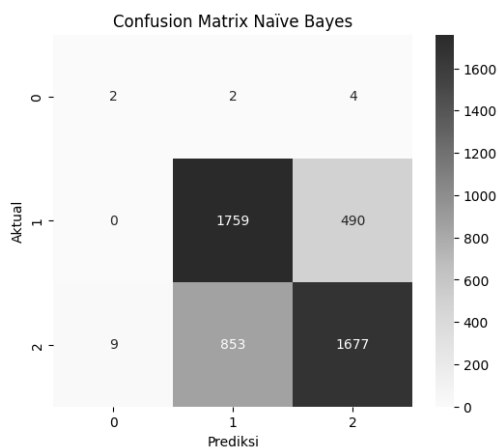
Berdasarkan hasil evaluasi, model menghasilkan akurasi sebesar 0,7168 (71,68%). Nilai ini menunjukkan bahwa secara keseluruhan model mampu memprediksi kategori magnitudo dengan tingkat ketepatan menengah. Tabel 1 menampilkan ringkasan performa model.

Tabel 1. Tabel Hasil Klasifikasi Gaussian Naïve Bayes

Kategori	Precision	Recall	F1-score	Support
Kuat	0.18	0.25	0.21	8
Ringan	0.67	0.78	0.72	2249
Sedang	0.77	0.66	0.71	2539
Accuracy			0.72	4796
Macro avg	0.54	0.56	0.55	4796
Weighted avg	0.72	0.72	0.72	4796

Precision tertinggi ada di kelas Sedang (0,77), tanda ketepatan prediksinya cukup baik. Recall tertinggi ada di kelas Ringan (0,78), artinya model ini cukup bagus untuk mendeteksi gempa ringan. Variasi performa seperti ini juga muncul di literatur lain. Debnath dkk. [8] menerapkan Bayes Net (varian klasifikasi berbasis Bayes) untuk memprediksi rentang magnitudo gempa di beberapa wilayah India dan mencatat akurasi yang jauh berbeda antar wilayah, dari 98,18% di Gujarat hingga performa yang jauh lebih rendah di wilayah lain dengan karakteristik kelas yang lebih menantang. Ini menguatkan pola yang sama: performa classifier berbasis Bayes sangat bergantung pada karakteristik distribusi data yang dipakai, bukan cuma pada algoritmanya. Bandingkan dengan Ariadi [6], yang mencatat akurasi Naïve Bayes 89,52% pada dataset kredit terstruktur. Selisih ini

menunjukkan bahwa performa Naïve Bayes sangat bergantung pada karakteristik distribusi datanya sendiri. Pada data seismik yang timpang seperti dalam penelitian ini, wajar kalau performanya lebih rendah. Kelas Kuat mencatat F1-score 0,21, dengan hanya 2 dari 8 sampel uji terdeteksi benar. Angka ini kecil, tapi Gaussian Naïve Bayes tetap jadi satu-satunya algoritma dalam penelitian ini yang berhasil mendeteksi kelas minoritas tersebut sama sekali. SMOTE [12] dan *cost-sensitive learning* layak dipertimbangkan untuk penelitian selanjutnya.



Gambar 6. *Confusion Matrix Naïve Bayes*

Cross-Validation dan Perbandingan Algoritma

Lima fold cross-validation menghasilkan rata-rata akurasi 68,78% ($\pm 7,32\%$), dengan skor per fold: 66,89%, 75,81%, 74,71%, 55,55%, dan 70,93%. Variasi antar fold cukup besar. Penyebabnya bisa ditebak, yaitu kelas Kuat yang jumlahnya kecil membuat proporsinya berubah-ubah drastis dari satu fold ke fold lain. Tabel 2 membandingkan performa ketiga algoritma pada dataset yang sama.

Tabel 2. Perbandingan Performa Algoritma Klasifikasi

Algoritma	Akurasi	Kuat (F1)	Ringan (F1)	Sedang (F1)	CV 5-Fold
Gaussian Naïve Bayes	71,68%	0,21	0,72	0,71	68,78% ($\pm 7,32\%$)
Random Forest	83,63%	0,00*	0,83	0,85	n/a
SVM (RBF)	79,15%	0,00*	0,78	0,80	n/a

*Random Forest dan SVM tidak berhasil mendeteksi kelas Kuat sama sekali.

Dari segi akurasi keseluruhan, Random Forest (83,63%) dan SVM (79,15%) sama-sama unggul dibanding Naïve Bayes. Tapi keduanya nihil di kelas Kuat, F1-nya 0,00. Naïve Bayes justru satu-satunya yang berhasil menangkap sebagian kelas Kuat ini, kelas yang paling menentukan dalam konteks mitigasi bencana [2]. Fernández-Gómez dkk. [9] menghadapi masalah yang sama persis saat memprediksi gempa bermagnitudo besar di Chili: classifier konvensional gagal menangani data yang timpang karena gempa besar selalu jadi kelas minoritas, sehingga mereka harus memakai teknik klasifikasi khusus data tidak seimbang dan *ensemble learning* untuk bisa mendeteksinya sama sekali. Ini menegaskan satu hal sederhana: akurasi keseluruhan yang tinggi tidak ada artinya kalau kelas paling berbahaya justru luput dari deteksi. Sebagai pembandingan lain, Pratiwi dan Arum [13] menerapkan Random Forest pada data gempa BMKG Indonesia dan mencatat akurasi 97% pada dataset yang lebih kecil dan lebih seimbang. Ini menunjukkan bahwa algoritma *ensemble* punya potensi besar kalau data yang dipakai lebih representatif. Untuk sekarang, model Naïve Bayes dalam penelitian ini cocok dipakai sebagai alat skrining awal berbasis data guna mendukung analisis BMKG, tapi belum cukup andal untuk keputusan operasional langsung seperti aktivasi prosedur evakuasi atau peringatan publik.

KESIMPULAN

Penelitian ini menerapkan Gaussian Naïve Bayes untuk mengklasifikasikan magnitudo gempa bumi di Indonesia dengan data BMKG periode 2023-2025. Setelah pra-pemrosesan, tersisa 23.979 data valid, terbagi menjadi 19.183 data latih dan 4.796 data uji. Model ini mencapai akurasi 71,68%, dengan F1-score 0,72 di kelas Ringan dan 0,71 di kelas Sedang. Di kelas Kuat, precision-nya 0,18 dan recall-nya 0,25, kecil memang, tapi cukup untuk menjadikan Naïve Bayes satu-satunya algoritma yang mendeteksi kelas minoritas ini. Random Forest (83,63%) dan SVM (79,15%) sama sekali gagal di kelas ini. Lima fold cross-validation menghasilkan rata-rata akurasi 68,78% ($\pm 7,32\%$).

Penelitian ini punya tiga keterbatasan. Pertama, ketidakseimbangan kelas yang ekstrem (kelas Kuat cuma 0,16% dari total data) membuat model condong ke kelas mayoritas. Kedua, asumsi independensi fitur Naïve Bayes tidak sepenuhnya cocok dengan karakteristik data seismik yang saling berkorelasi. Ketiga, akurasi 71,68% masih jauh dari cukup untuk pengambilan keputusan operasional langsung. Karena itu, model ini paling pas diposisikan sebagai alat bantu skrining awal dalam sistem peringatan dini berbasis data, bukan pengganti keputusan akhir. Untuk penelitian selanjutnya, penerapan SMOTE untuk mengatasi *class imbalance*, eksplorasi XGBoost atau LightGBM sebagai metode *ensemble*, serta validasi model pada wilayah seismik lain di Indonesia layak jadi prioritas.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Pak Akhiyar Waladi, S.Komp., M.Kom., Ibu Ulfa Khaira, S.Komp., M.Kom., dan Pak Pradita Eko Prasetyo Utomo, S.Pd., M.Cs. atas bimbingan dan arahan yang diberikan selama penelitian ini berlangsung.

DAFTAR PUSTAKA

- [1] H. Afif *et al.*, "Local earthquake tomography of the source region of the 2018 Lombok earthquake sequence, Indonesia," *Geophys. J. Int.*, vol. 226, no. 3, pp. 1814–1823, 2021, doi: 10.1093/gji/ggab189.
- [2] R. Jena *et al.*, "Integrated model for earthquake risk assessment using neural network and analytic hierarchy process: Aceh province, Indonesia," *Geoscience Frontiers*, vol. 11, no. 2, pp. 613–634, 2020, doi: 10.1016/j.gsf.2019.07.006.

- [3] K. M. Asim, F. Martínez-Álvarez, A. Basit, and T. Iqbal, "Earthquake magnitude prediction in Hindukush region using machine learning techniques," *Natural Hazards*, vol. 85, no. 1, pp. 471–486, 2017, doi: 10.1007/s11069-016-2579-3.
- [4] D. P. Sari and M. Hariadi, "Klasifikasi Magnitudo Gempa Bumi di Pulau Jawa Menggunakan Decision Tree," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 3, pp. 521–528, 2022.
- [5] I. Maulita and A. M. Wahid, "Prediksi Magnitudo Gempa Menggunakan Random Forest, Support Vector Regression, XGBoost, LightGBM, dan Multi-Layer Perceptron Berdasarkan Data Kedalaman dan Geolokasi," *Jurnal Pendidikan Teknologi Indonesia*, 2024, doi: 10.52436/1.jpti.470.
- [6] F. Ariadi, "Analisa Perbandingan Algoritma DT C.45 dan Naïve Bayes Dalam Prediksi Penerimaan Kredit Motor," *KERNEL: Jurnal Riset Inovasi Bidang Informatika dan Pendidikan Informatika*, vol. 1, no. 1, pp. 1–8, 2020, doi: 10.31284/j.kernel.2020.v1i1.1183.
- [7] S. D. Y. Kusuma, H. Al Islam, and P. Rosyani, "Penerapan Naive Bayes Untuk Klasifikasi Penyakit Endokrin Pada Pasien Lansia," *KERNEL: Jurnal Riset Inovasi Bidang Informatika dan Pendidikan Informatika*, vol. 5, no. 2, pp. 1–10, 2024, doi: 10.31284/j.kernel.2024.v5i2.7312.
- [8] P. Debnath *et al.*, "Analysis of Earthquake Forecasting in India Using Supervised Machine Learning Classifiers," *Sustainability*, vol. 13, no. 2, p. 971, 2021, doi: 10.3390/su13020971.
- [9] M. J. Fernández-Gómez, G. Asencio-Cortés, A. Troncoso, and F. Martínez-Álvarez, "Large earthquake magnitude prediction in Chile with imbalanced classifiers and ensemble learning," *Applied Sciences*, vol. 7, no. 6, p. 625, 2017, doi: 10.3390/app7060625.
- [10] BMKG and USGS, "Earthquakes in Indonesia," 2025, *Kaggle*. doi: 10.34740/KAGGLE/DSV/13265963.
- [11] D. J. Hand and K. Yu, "Idiot's Bayes: Not so stupid after all?," *International Statistical Review*, vol. 69, no. 3, pp. 385–398, 2001.
- [12] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [13] A. P. Pratiwi and P. R. Arum, "Implementation of Random Forest Algorithm to Classify Earthquake in Indonesia," *Eigen Mathematics Journal*, vol. 8, no. 1, 2025, doi: 10.29303/emj.v8i1.185.