

“AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN
DE LA ECONOMÍA PERUANA”.



ESCUELA PROFESIONAL DE CIENCIA DE LA
COMPUTACIÓN
TÓPICOS CIENCIA DE DATOS

Análisis Visual Interactivo y
Exploración de Patrones
Espacio-Temporales en Eventos
Sísmicos del USGS

Estudiante:

Andrea del Rosario Lopez Condori

Docente :

ANA MARIA CUADROS
VALDIVIA



Introducción

La correcta interpretación de los eventos sísmicos es un componente clave para la gestión del riesgo de desastres naturales. Sin embargo, los registros sísmicos públicos, como los del United States Geological Survey (USGS), contienen tanto terremotos naturales como eventos artificiales —explosiones de cantera, colapsos, pruebas inducidas— que, aunque físicamente diferentes en origen, pueden compartir características similares en magnitud, profundidad o localización. Esta mezcla no diferenciada complica los sistemas de alerta, los estudios de peligro sísmico y las decisiones de protección civil. Por lo tanto, el tratamiento y análisis de estos datos no es solo un problema técnico, sino también un problema social con implicaciones reales para la seguridad pública y la planificación urbana.

Diversos trabajos han propuesto soluciones computacionales para abordar esta situación. Kassa y Dugda (1) aplicaron modelos supervisados (Random Forest, SVM) junto a visualización exploratoria para clasificar eventos sísmicos, destacando el uso de mapas y análisis de correlación. El-Hadidy et al. (2) demostraron que con solo tres variables —frecuencia de esquina, potencia y razón espectral— se puede lograr una clasificación binaria con más del 99 % de precisión, utilizando árboles de decisión avanzados. Por otro lado, Hajibabae et al. (3) aplicaron técnicas de reducción de dimensionalidad (t-SNE) y balanceo de clases (SMOTE) para representar visualmente agrupamientos y patrones estructurales en datos sísmicos, lo cual facilitó la diferenciación entre tipos de eventos sin depender de etiquetas explícitas.

No obstante, muchas de estas soluciones están enfocadas en la clasificación automática como fin principal, dependen de modelos supervisados con grandes volúmenes de datos etiquetados o requieren características derivadas difíciles de generalizar. Además, la etapa de análisis visual a menudo se trata de forma superficial, a pesar de que es crucial para comprender la naturaleza del dataset y sus limitaciones. En especial, en catálogos abiertos como el del USGS, los datos presentan desafíos particulares: desbalance de clases, solapamiento de variables como magnitud y profundidad, y distribución espacial desigual.

En respuesta a esta situación, el presente proyecto propone un análisis visual detallado de eventos sísmicos naturales y artificiales usando datos tabulares del USGS comprendidos entre 1990 y 2024. A través de visualizaciones como mapas de calor, boxplots, histogramas y gráficos de dispersión, se busca explorar la estructura y comportamiento de las variables sísmicas. Como complemento exploratorio, se integran técnicas de agrupamiento no supervisado (como K-Means) para identificar agrupamientos o separaciones emergentes en los datos, sin que estas dependan de etiquetas previas.

Este enfoque se sitúa en la frontera entre la visualización, la estadística y la computación, y traduce un problema social —la necesidad de comprender mejor los eventos sísmicos registrados— en una solución computacional basada en el análisis visual de datos, con modelos como herramienta de apoyo y no como fin.

Trabajos relacionados

Diversos estudios recientes han abordado la clasificación automática de eventos sísmicos a partir de datos estructurados, aplicando modelos de aprendizaje automático y técnicas de visualización. A continuación, se describen diez trabajos representativos que abordan enfoques similares al del presente proyecto.

Kassa y Dugda (1) propusieron el uso de múltiples clasificadores supervisados, como Random Forest, SVM y KNN, para categorizar eventos sísmicos en múltiples clases utilizando variables como magnitud, profundidad y coordenadas geográficas. Su enfoque incluyó visualizaciones exploratorias como mapas de dispersión y matrices de correlación. Obtuvieron una precisión del 93.5 % con Random Forest, superando ampliamente a SVM, que alcanzó solo un 34.7 %.

El-Hadidy et al. (2) desarrollaron un modelo de clasificación binaria para distinguir entre terremotos y explosiones de cantera. Aplicaron técnicas de selección de atributos e importancia de variables, y lograron una precisión del 99.68 % utilizando únicamente tres variables: frecuencia de esquina, potencia del evento ($\log(P_e)$) y razón espectral.

Hajibabae et al. (3) utilizaron t-SNE para la reducción de dimensionalidad y visualización de datos sísmicos estructurados, en combinación con técnicas de balanceo como SMOTE. Su análisis permitió identificar patrones visuales entre clases, demostrando que incluso con alta dimensionalidad, es posible obtener separabilidad mediante proyecciones 2D.

Malfante et al. (5) diseñaron un sistema de clasificación de señales sísmicas volcánicas mediante la extracción de más de 100 descriptores físicos y el uso de modelos como SVM y Random Forest. Reportaron precisiones superiores al 90 % al identificar diferentes tipos de eventos volcánicos, lo que demuestra la aplicabilidad de modelos clásicos con ingeniería de características avanzada.

Arul y Kareem (6) emplearon transformación de Shapelets sobre series temporales sísmicas, de viento y oleaje. Su enfoque mejoró la interpretabilidad del modelo y logró precisiones superiores al 85 % en datasets multivariados, contribuyendo a una mejor comprensión de la estructura temporal de los datos.

Un estudio anónimo publicado en *Applied Geophysics* (4) aplicó modelos tradicionales de clasificación como SVM y Random Forest sobre un dataset tabular con información de magnitud y profundidad. Los resultados destacaron diferencias claras entre clases y se apoyaron en visualizaciones como histogramas y matrices de confusión.

Un trabajo reciente en la revista *Geomechanics and Geophysics for Geo-Energy and Geo-Resources* (7) propuso optimizar la clasificación de facies sísmicas mediante la selección automática de atributos utilizando algoritmos como ReliefF, SFS y Random Forest. Los autores lograron reducir la complejidad del modelo manteniendo una alta precisión.

Investigadores de NHESS (8) desarrollaron un sistema automático en tiempo real para clasificar señales sísmicas asociadas a deslizamientos de tierra. Utilizando Random Forest, lograron altos niveles de recall y eficiencia, con aplicaciones prácticas en monitoreo continuo.

Münchmeyer et al. (9) realizaron una evaluación cuantitativa de algoritmos de aprendizaje profundo para la identificación de fases sísmicas en señales crudas. Aunque sus resultados superaron a métodos tradicionales en precisión temporal, su enfoque no es aplicable directamente a datos tabulares como los del USGS, por requerir acceso a formas de onda.

Finalmente, un estudio presentado en una conferencia técnica (10) empleó modelos clásicos de clasificación como SVM y KNN para distinguir entre eventos naturales y artificiales usando datos tabulares. Los autores concluyeron que la precisión mejora significativamente cuando se incluyen atributos como profundidad y energía estimada.

Estos trabajos demuestran que la combinación de modelos supervisados, selección de variables y visualización exploratoria ofrece un enfoque eficaz para la clasificación

automática de eventos sísmicos, alineado con los objetivos del presente proyecto.

1. Metodología

La presente investigación sigue un enfoque centrado en el análisis visual y exploratorio de datos, alineado con los principios de la Ciencia de Datos. En lugar de enfocarse en modelos supervisados para clasificación automática, se prioriza la comprensión profunda de los eventos sísmicos mediante visualizaciones interactivas y agrupamiento no supervisado como herramienta de apoyo. A continuación, se detallan las etapas seguidas:

1. **Recolección de datos:** Se utilizaron registros tabulares del catálogo público del *United States Geological Survey* (USGS), comprendidos entre los años 1990 y 2024. Las variables consideradas incluyen magnitud, profundidad, latitud, longitud, tipo de evento, fecha y hora.
2. **Preprocesamiento de datos:** Se aplicaron técnicas de limpieza para eliminar registros incompletos, se codificaron variables categóricas (como el tipo de evento), y se estandarizaron las variables numéricas. Esto permitió preparar los datos para su análisis visual posterior.
3. **Análisis exploratorio (EDA):** Se calcularon estadísticas descriptivas para identificar distribuciones, valores atípicos y relaciones generales entre las variables. Esta etapa permitió conocer el comportamiento global del dataset.
4. **Análisis visual interactivo:** Se desarrollaron visualizaciones como mapas geográficos, histogramas, boxplots, gráficos de dispersión y series temporales. Estas representaciones gráficas ayudaron a identificar patrones diferenciadores entre eventos naturales y artificiales sin necesidad de etiquetas supervisadas.
5. **Agrupamiento no supervisado:** Como complemento al análisis visual, se aplicaron algoritmos como *K-Means* y *t-SNE*, con el fin de detectar agrupamientos emergentes en los datos. Estas técnicas ayudaron a validar visualmente la posible separación entre eventos sísmicos de diferente origen.
6. **Interpretación y conclusiones:** Finalmente, se sintetizaron los hallazgos visuales y computacionales. Se identificaron las variables más relevantes para diferenciar eventos, se evaluó la utilidad de las técnicas aplicadas y se propusieron recomendaciones para futuros estudios similares.

A continuación, se presenta el flujo metodológico seguido en el proyecto:

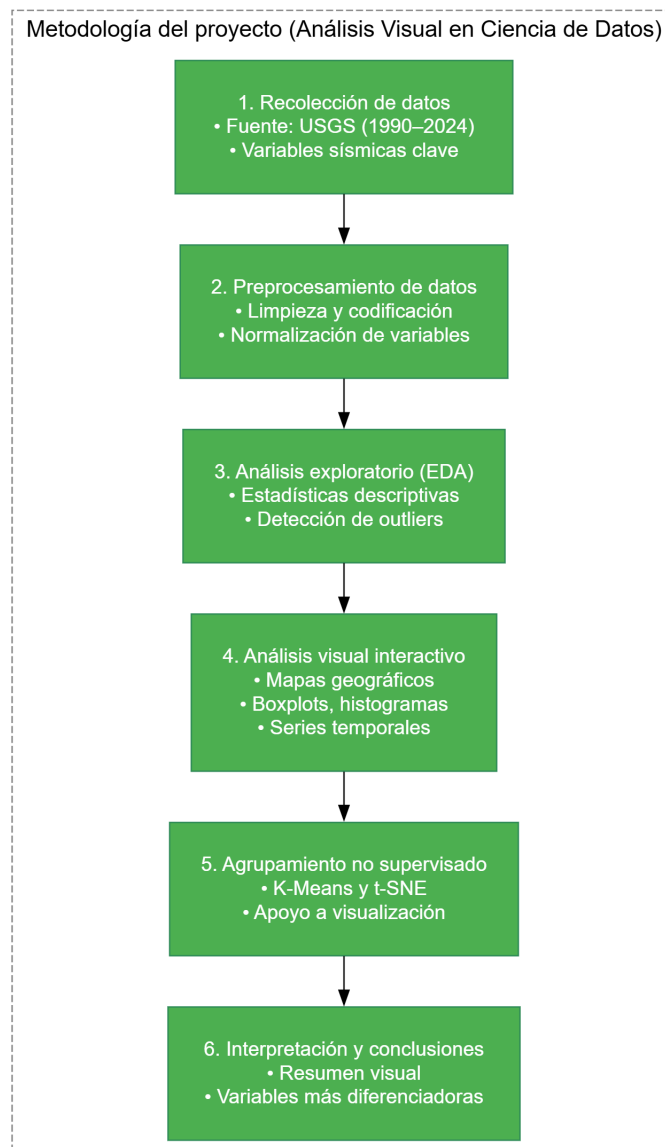


Figura 1: Proposed Method

Referencias

- [1] Kassa, A. B. & Dugda, M. T. (2025). *Implementation of Machine Learning Algorithms for Seismic Events Classification*. arXiv preprint arXiv:2502.05197.
- [2] El-Hadidy, M., et al. (2024). *Enhancing the classification of seismic events with supervised machine learning and feature importance*. Scientific Reports, 14(1). <https://doi.org/10.1038/s41598-024-81113-7>
- [3] Hajibabaei, P., Pourkamali-Anaraki, F., Hariri-Ardebili, M. A. (2021). *An Empirical Evaluation of the t-SNE Algorithm for Data Visualization in Structural Engineering*. arXiv preprint arXiv:2109.08795.
- [4] Anónimo. (2021). *Analysis of Seismic Data Using Machine Learning Algorithms*. Journal of Applied Geophysics.

- [5] Malfante, M., Hibert, C., Maggi, A., Provost, F., et al. (2018). *Automatic Classification of Volcano Seismic Signatures*. Journal of Geophysical Research: Solid Earth.
- [6] Arul, M. & Kareem, A. (2020). *Applications of Shapelet Transform to Time Series Classification of Earthquake, Wind and Wave Data*. arXiv preprint arXiv:2004.11243.
- [7] Autor, Varios (2024). *Enhancing machine learning-based seismic facies classification through attribute selection*. Geomechanics and Geophysics for Geo-Energy and Geo-Resources.
- [8] Autor, Varios (2021). *Near-real-time automated classification of seismic signals of slope failures with continuous random forests*. Natural Hazards and Earth System Sciences.
- [9] Münchmeyer, J., Woollam, J., Rietbrock, A., Tilmann, F., et al. (2021). *Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers*. arXiv preprint arXiv:2110.13671.
- [10] Anónimo (2021). *Seismic Data Classification Using Machine Learning*. ResearchGate / IOP Conf. Series.