

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

**Clasificación Automática de Eventos Sísmicos basado en
Wrapper y Deep Learning aplicado a registros
acelerográficos**

Mateo Ariel Moreno Peñaloza

Ingeniería Civil

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniería Civil

Quito, 17 de diciembre de 2022

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

**HOJA DE CALIFICACIÓN
DE TRABAJO DEL AVANCE 3**

**Clasificación Automática de Eventos Sísmicos basado en Wrapper y Deep
Learning aplicado a registros acelerográficos**

Mateo Ariel Moreno Peñaloza

Fabricio Yepez, Ph.D.

Quito, 17 de diciembre de 2024

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: Mateo Ariel Moreno Peñaloza

Código: 00215814

Cédula de identidad: 1804428751

Lugar y fecha: Quito, 16 de diciembre de 2024

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017). Discussion document on best practice for issues around theses publishing, disponible en <http://bit.ly/COPETheses>.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on <http://bit.ly/COPETheses>.

RESUMEN

La clasificación de eventos sísmicos es una necesidad creciente debido a los beneficios que ofrece a la sociedad, como los sistemas de alerta temprana, la optimización de la respuesta ante emergencias y la resiliencia de infraestructuras. Las herramientas de Machine Learning brindan un enfoque práctico para la predicción sísmica, contribuyendo así a los sistemas de alerta temprana. Este estudio se centra en desarrollar un método automático de clasificación multiclase utilizando estrategias wrapper basadas en metaheurísticas y clasificadores de shallow learning para maximizar la clasificación del mecanismo focal en datos de movimiento sísmico, es decir, normal-oblique, reverse-oblique o strike-slip. El módulo wrapper implementó dos algoritmos metaheurísticos en combinación con tres clasificadores de Machine Learning como funciones objetivo para evaluar cada subconjunto de características en términos del área bajo la curva característica operativa del receptor (ROC). Posteriormente, el módulo de clasificación utilizó cuatro clasificadores distintos con los subconjuntos de características para determinar el mejor esquema de clasificación entre varias soluciones candidatas. El método propuesto fue entrenado y validado en un conjunto de datos de movimiento sísmico con 465 vectores de características. Los resultados destacaron el subconjunto de 9 características proporcionado por el algoritmo particle swarm optimization en combinación con el clasificador naive bayes como el subconjunto que de mejor manera reduce la cantidad de características del sistema y mantiene los errores de clasificación del mecanismo focal al mínimo, alcanzando una puntuación AUC de 0.79. Este resultado prometedor constituye un paso favorable hacia la mejora de la clasificación automática del movimiento sísmico y su implementación en sistemas de alerta temprana.

Palabras clave: Clasificación de terremotos, Modelo wrapper, Algoritmos metaheurísticos, Clasificadores machine learning.

ABSTRACT

Earthquake classification is a growing necessity due to the numerous benefits it provides to society, such as early warning systems, emergency response optimization and infrastructure resilience. Machine Learning tools offer a practical approach to seismic prediction and, thus, to further early warning systems. This study focuses on developing an automatic multiclass classification method using metaheuristic-based wrapper strategies and shallow learning classifiers to maximize the primary focal mechanism classification on seismic motion data, i.e., normal-oblique, reverse-oblique, or strike-slip. The wrapper module implemented two metaheuristic algorithms in combination with three machine learning classifiers as objective functions to measure each feature subset in terms of area under the receiver operating characteristic curve scores. Then, the classification module feeds four different machine learning classifiers with the feature subsets to determine the best classification scheme among several candidate solutions. The proposed method was trained and validated on a seismic motion dataset with 465 feature vectors. The results highlighted the subset of 9 features provided by the Particle Swarm Optimization algorithm in combination with the Naive Bayes classifier as the subset that most effectively reduces the system's feature count while keeping focal mechanism classification errors to a minimum, achieving an AUC score of 0.79. This promising result constitutes a favorable step toward improving the automatic classification of seismic motion and its implementation into early warning systems.

Keywords: Earthquake classification, Wrapper model, Metaheuristic algorithms, Machine learning classifiers.

TABLA DE CONTENIDO

Introducción	10
Desarrollo.....	16
Base de datos.....	16
Método Propuesto	16
Procesamiento de señales y creación de conjuntos de datos experimentales	18
Preprocesamiento de datos.....	18
Configuración de propiedades metaheurísticas y MLCs	19
Criterio de selección	21
Software	21
Resultados	22
Rendimiento de la fase del método propuesto	23
Mejor resultado del método propuesto	24
Relevancia de características	25
Conclusiones	28
Referencias bibliográficas.....	29

ÍNDICE DE TABLAS

Tabla #1: Resultados de desempeño del método propuesto.....	22
Tabla #2: Porcentaje de aciertos de eventos sísmicos.....	24
Tabla #3: Evaluación de rendimiento.....	26

ÍNDICE DE FIGURAS

Figura #1: Diagrama del método propuesto.....	17
Figura #2: Procesamiento de datos (Extraído de Weka).....	19
Figura #3: Parámetros Genetic Algorithm (Extraído de Weka).....	19
Figura #4: Parámetros Particle Swarm Optimisation (Extraído de Weka).....	20
Figura #5: Parámetros de los MLC (Extraído de Weka).....	21
Figura #6: Matriz de confusión PSO+NB+kNN (k=5).....	24
Figura #7: Matriz de confusión GA+RF+kNN (k=3).....	25
Figura #8: Forma 2D del mecanismo focal SS.....	26
Figura #9: Forma 2D del mecanismo focal NO.....	26
Figura #10: Forma 2D del mecanismo focal RO.....	27

INTRODUCCIÓN

Los terremotos son eventos resultantes del movimiento de la corteza terrestre debido a la liberación de energía en las placas tectónicas. Ecuador, ubicado en el Anillo de Fuego del Pacífico, es especialmente vulnerable a estos fenómenos. Históricamente, el país ha experimentado ciclos sísmicos aproximadamente cada 50 años, como los ocurridos entre 1900-1910, 1940-1950, y a principios del siglo XXI. Eventos recientes, como el devastador terremoto de Pedernales en 2016, subrayan la urgencia de desarrollar herramientas de protección sísmica para mitigar los efectos de futuros desastres naturales (Theurer, Jiménez, Velasco y Zambrano, 2017).

La imprevisibilidad de los terremotos ha causado innumerables tragedias, destacando la necesidad de herramientas de predicción temprana efectivas (Lomnitz, 2013). Aunque pronosticarlos no es posible, obtener información precisa y rápida sobre los sismos resulta crucial para guiar las acciones inmediatas. Actualmente, los sismólogos emplean ondas de compresión y extensión medidas por sismógrafos para determinar la ubicación y magnitud de los terremotos. Sin embargo, definir mecanismos focales aún requiere intervención humana, lo que limita la automatización y retrasa la respuesta.

Métodos tradicionales como las Redes sísmicas en tiempo real de California (Allen & Melgar, 2019) y los Modelos de ciclo sísmico (Scholz, 2019) han sido útiles, pero enfrentan limitaciones como dependencia de datos restringidos, su enfoque lineal y la falta de aprendizaje continuo. Ante estos desafíos, el Machine Learning (ML) surge como una alternativa prometedora. A diferencia de los métodos convencionales, ML procesa grandes volúmenes de datos, identifica patrones complejos y mejora constantemente mediante aprendizaje iterativo.

Este enfoque no solo establece nuevas bases para el análisis de terremotos, sino que también puede salvar vidas y reducir daños en infraestructura mediante alertas tempranas. Así, ML se convierte en una herramienta clave para la gestión de desastres naturales.

Dado que este estudio emplea herramientas avanzadas de aprendizaje automático para la clasificación de terremotos, es importante definir ciertos términos clave que serán utilizados a lo largo del trabajo. Esto facilitará una comprensión más clara de los métodos y enfoques propuestos.

1. Mecanismos Focales Sísmicos (Strike-Slip, Reverse Oblique, Normal Oblique)

Los mecanismos focales describen la orientación y el movimiento de las fallas sísmicas en el origen de un terremoto. Los tres tipos principales son: strike-slip (deslizamiento lateral), reverse oblique (oblicuo inverso) y normal oblique (oblicuo normal). Estas categorías dependen de la dirección del esfuerzo tectónico aplicado en la falla (Aki & Richards, 2002).

2. Machine Learning (ML)

Machine learning o aprendizaje automático, es un campo que permite a las computadoras aprender patrones en datos sin necesidad de ser programadas explícitamente para cada tarea (Mitchell, 1997). En sismología, ML analiza señales sísmicas con precisión y eficiencia (Domingos, 2012). En el contexto de esta investigación, ML es una herramienta clave para analizar señales sísmicas y clasificar eventos de manera precisa y eficiente.

3. Shallow Learning

Rama de la computación que permite a las máquinas aprender patrones a partir de datos sin necesidad de programación explícita para cada tarea (Goodfellow, Bengio, &

Courville, 2016). Este tipo de aprendizaje es especialmente adecuado para problemas de clasificación, ya que optimiza los recursos computacionales sin sacrificar la precisión en tareas específicas.

4. Wrapper

Los métodos wrapper son técnicas de selección de características que buscan optimizar el rendimiento de un modelo a partir de la combinación de subconjuntos de características.

Los wrappers son eficaces en machine learning porque permiten adaptar el espacio de características de acuerdo con las necesidades específicas del modelo y los datos disponibles (Guyon & Elisseeff, 2003).

5. Algoritmos Metaheurísticos

En el contexto de este estudio, los algoritmos metaheurísticos funcionan como filtros que agilizan el proceso de clasificación, asegurando un equilibrio entre precisión y eficiencia.

En este estudio se utilizan dos algoritmos metaheurísticos muy populares que son:

- a. Genetic Algorithm (GA): Se inspira en el proceso de selección natural, aplicando los principios de selección, cruce y mutación a las características para encontrar la mejor solución en cada generación (McCall, 2005).
- b. Particle Swarm Optimization (PSO): Se inspira en los comportamientos sociales de los grupos de aves y los bancos de peces. Su comportamiento está regido por tres procesos principales: velocidad, actualización de posición y movimiento iterativo, aplicados a las características para encontrar la mejor solución de cada enjambre (Ghamisi & Benediktsson, 2014).

6. Machine Learning Classifier (MLC)

Un machine learning classifier es un modelo que procesa y clasifica datos basándose en patrones identificados durante el entrenamiento. En el análisis de señales sísmicas, los MLC permiten determinar qué tan efectivos son los atributos, de esta manera se conoce la relevancia de cada característica de un espacio de datos y así seleccionar a los mejores.

En este estudio se utilizó un total de 7 MLCs de diferentes taxonomías que son:

- a. Naved Bayes (NB): Es un clasificador probabilístico simple basado en el teorema de Bayes, que asume independencia entre los predictores (Kupervasser, 2014).
- b. Random Forest (RF): Es un algoritmo de aprendizaje de conjuntos que construye una multitud de árboles de decisión y determina la mejor clasificación final mediante votación mayoritaria (Zwane, Tarwireyi, & Adigun, 2018)
- c. Linear Discriminant Análisis (LDA): Es un clasificador que proyecta características en un espacio lineal para maximizar la separabilidad de clases (Xanthopoulos et al., 2013)
- d. Quadratic Discriminant Analysis (QDA): Una variante de LDA que permite fronteras de clase no lineales. Su eficiencia computacional es $O(n^2)$ (Naveen & Diwan, 2020).
- e. k-Nearest Neighbors (kNN): Es un clasificador perezoso donde la instancia desconocida se clasifica según el voto mayoritario de sus k vecinos más cercanos. Su eficiencia computacional es $O(n)$ (Li et al., 2020).
- f. Multilayer Perceptron (MLP): Una clase de redes neuronales artificiales feedforward que proporciona un mapeo no lineal entre entradas y salidas mientras minimiza el error mediante back-propagation. Su eficiencia computacional es $O(n)$ (Zwane et al., 2018).

- g. Support Vector Machine (SVM): Un clasificador que encuentra el hiperplano óptimo en un espacio de alta dimensión para separar clases. Es especialmente efectivo cuando los datos no son linealmente separables, utilizando funciones de núcleo. Su eficiencia computacional es $O(n)$ (Zwane et al., 2018).

A la eficiencia computacional mencionada es conocida con BigO, esta hace referencia a una manera de describir qué tan rápido crece el tiempo o los recursos necesarios para resolver un problema a medida que aumentas el tamaño de los datos de entrada. Por lo ejemplo un $O(n)$ demuestra que el tiempo de trabajo aumenta de manera lineal y $O(n^2)$ que el tiempo de trabajo aumenta de manera cuadrática.

7. Enfoque híbrido

El enfoque híbrido se refiere a la combinación de algoritmo y clasificadores de shallow learning para reducir el espacio de características, maximizando la precisión de clasificación al mismo tiempo que se minimizan los costos computacionales.

8. Método de validación cruzada

Es un método estadístico utilizado para evaluar el rendimiento de un modelo de clasificación, dividiendo los datos en varios subconjuntos o folds. En este proceso, el modelo se entrena en algunos de estos subconjuntos y se prueba en los restantes, repitiendo este ciclo varias veces para obtener una evaluación robusta del modelo. Este enfoque es crucial en la clasificación de eventos sísmicos, ya que garantiza que el modelo no se ajuste a un subconjunto específico de datos, lo que podría llevar a resultados sesgados. (James, Witten, Hastie, & Tibshirani, 2013).

Luego de establecer los conceptos clave relacionados con la clasificación y análisis de señales sísmicas, es más sencillo el profundizar en la metodología de este estudio. El uso de ML en el

procesamiento de datos sísmicos no es nuevo, investigaciones previas en este campo, se han enfocado principalmente en la predicción de terremotos (Salam, Ibrahim, & Abdelminaam, 2021) y la clasificación de terremotos (Bayliss et al., 2020). Aunque ambas vías de investigación son de gran importancia en la evolución de la predicción de eventos sísmicos, este estudio se concentra específicamente en la clasificación de terremotos. Estudios enfocados en esta área destacan que los MLCs y enfoques híbridos son una de las estrategias más prometedoras en torno a la clasificación de terremotos. Estos métodos se basan en la reducción del espacio de características de una compleja base datos, mediante algoritmos que permiten a los clasificadores maximizar el rendimiento en conjuntos completos o reducidos de atributos.

Por lo que, se propone un método automático de clasificación multiclase utilizando estrategias wrapper basadas en metaheurísticas y clasificadores de shallow learning para maximizar la clasificación del mecanismo focal primario en datos de movimiento sísmico. La contribución detrás de este objetivo es reducir el espacio de características original desde un punto de vista bioinspirado, mientras se maximiza el rendimiento de clasificación en tres clases de actividad sísmica: strike-slip (SS), reverse-oblique (RO) y normal-oblique (NO). Esta novedosa combinación de algoritmos metaheurísticos basado en filtros más clasificadores shallow learning, proporciona una reducción de espacio de características más eficiente al abordar el riesgo de sobreajuste y el costo computacional que suelen acompañar a los métodos basados en wrapper.

En este estudio, se examina el uso de métodos de ML para la clasificación de terremotos, contribuyendo al avance sobre la investigación sísmica junto al uso de herramientas tecnológicas avanzadas como es ML.

DESARROLLO

Base de datos

La base de datos utilizada en este estudio proviene de PEER NGA-West2 (Ancheta et al., 2013) y contiene registros de movimientos sísmicos generados por terremotos de corteza somera en regiones tectónicas activas alrededor del mundo (Yépez et al., 2024). Incluye 599 eventos sísmicos distribuidos en 21336 datos en las 3 componentes con magnitudes de entre 3.0 y 7.9, distancias epicentrales de hasta 1500 km, y valores de V_{s30} que oscilan entre 94 y 2100 m/s. Los datos abarcan regiones como el oeste de EE. UU., Irán, China, Italia, México y Japón, e incluyen los principales mecanismos focales: Strike-Slip (SS), Reverse Oblique (RO) y Normal Oblique (NO).

Cada registro contiene información de ondas sísmicas captadas por un acelerómetro triaxial, que mide el movimiento vertical y los movimientos horizontales ortogonales x e y (Mousavi et al., 2019). En este análisis, sin embargo, se emplearon únicamente las componentes verticales, esto debido a que dichos datos son más fáciles de interpretar pues son los que mayor liberación de energía generan.

Método Propuesto

El estudio propone un método automatizado para la clasificación de terremotos según su mecanismo focal, basado en la selección de características. Este enfoque consta de dos pasos principales:

1. Selección inicial de características

El método utiliza mecanismos wrapper que combinan algoritmos metaheurísticos (GA y PSO) con clasificadores de aprendizaje automático (MLCs). Los wrappers generan

subconjuntos candidatos evaluados con la métrica de Área Bajo la Curva (AUC) para identificar los más prometedores (Shatnawi, 2024).

2. Validación cruzada y evaluación final

Los mejores subconjuntos son validados mediante un método de diez pliegues (Fontanari, Fróes, & Recamonde-Mendoza, 2022), empleando cuatro clasificadores adicionales para asegurar que los resultados no estén sesgados y seleccionar el esquema de clasificación óptimo, tal como se muestra en la Figura 1.

En el proceso de optimizar de manera efectiva la selección de características, los algoritmos GA y PSO buscan reducir la dimensionalidad de los datos, seleccionando únicamente las características más relevantes. Para evaluar los subconjuntos generados, se utilizaron tres clasificadores iniciales (NB, RF y LDA). Posteriormente, los mejores subconjuntos fueron evaluados con otros cuatro clasificadores para confirmar la robustez de los resultados, los cuatro MLCs referidos son: QDA, kNN, MLP y SVM.

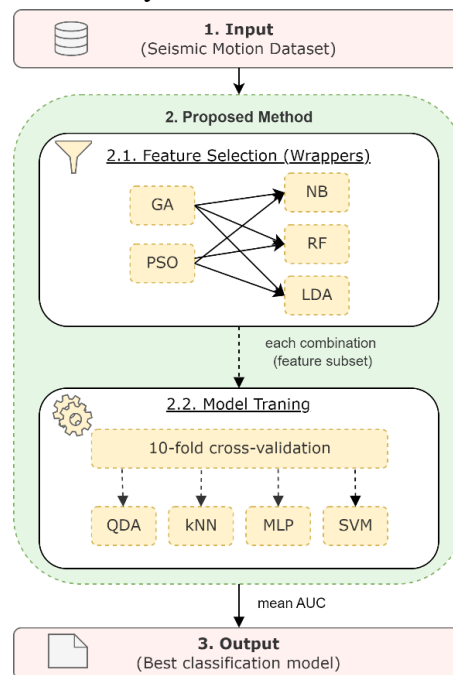


Figura #1: Diagrama del método propuesto

Procesamiento de señales y creación de conjuntos de datos experimentales

El procesamiento de las señales sísmicas en este estudio se basó en metodologías previamente desarrolladas, obteniendo características que personifican los datos sísmicos y segmentándolos en 186 eventos del tipo SS, 95 eventos del tipo NO y 184 eventos del tipo RO (Yépez, Benítez, Pérez, & Grijalva, 2024). Este proceso se dividió en tres etapas principales:

1. Cálculo del espectrograma

Se aplicó la Transformada de Fourier de Tiempo Corto (STFT) para convertir las señales del dominio del tiempo al de la frecuencia. Este enfoque maximizó la energía de las señales clave y minimizó pérdidas en los bordes.

2. Modelo de segmentación

Se eliminaron componentes de baja frecuencia, como el ruido, utilizando un umbral de potencia de 0 dB. Después, se aplicó el método de Otsu para determinar el umbral óptimo y el algoritmo de rastreo de vecinos de Moore para segmentar y visualizar los eventos sísmicos con claridad.

3. Cálculo y selección de características

Se calcularon 25 características agrupadas en tres categorías: estadísticas de intensidad, descriptores de forma y características de textura. Estas se seleccionaron con base en su relevancia demostrada para la clasificación de eventos sísmicos en estudios previos.

Preprocesamiento de datos

Los datos se escalaron al rango de 0 a 1 mediante normalización min-max, reduciendo la dispersión y mejorando el rendimiento de los clasificadores. El conjunto se dividió en una proporción de 90% para entrenamiento y 10% para pruebas, garantizando una evaluación fiable.

Posteriormente, los subconjuntos reducidos pasaron por un método de validación cruzada con diez iteraciones, generando varios conjuntos experimentales.

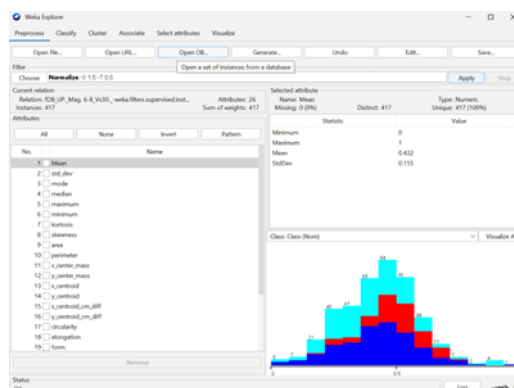


Figura #2: Procesamiento de datos (Extraído de Weka)

Configuración de propiedades metaheurísticas y MLCs

Los wrappers que utilizaron el algoritmo GA, se implementaron en una configuración basada en un estudio que considera a dicho método apto para seleccionar características relevantes a través de la clasificación (Meenachi & Ramakrishnan, 2021). Con base en dicha investigación, se estableció un tamaño de población inicial (pop) de 13 características, una longitud de cromosoma (cl) de 25 características (la dimensión de la base de datos), un rango de mutación (mr) de 0.2 unidades, y el número de padres (np) en la operación de cruce se fijó en 10 unidades. Además, el algoritmo fue configurado para ejecutarse durante un total de 500 iteraciones, excepto para el clasificador RF, donde las iteraciones de GA se limitaron a 200 debido al costo computacional asociado con cada bosque.

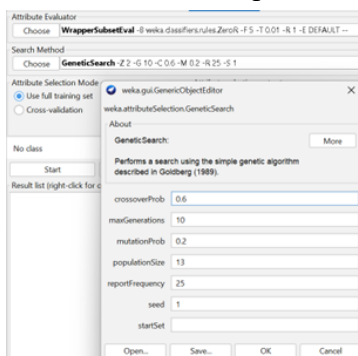


Figura #3: Parámetros Genetic Algorithm (Extraído de Weka)

Para los wrappers que utilizaron PSO, se implementó una configuración basada en un estudio que considera a dicho método apto como un método para seleccionar características relevantes a través de la clasificación (Xue, Zhang, & Browne, 2012). Con base en dicha investigación, se estableció un tamaño de población inicial (pop) de 13 características, una longitud de cromosoma (cl) de 25 características (la dimensión de la base de datos), constantes de aceleración c_1 y c_2 de 0.402, y una inercia de 0.196. Además, el algoritmo se configuró para ejecutarse durante un total de 100 iteraciones.

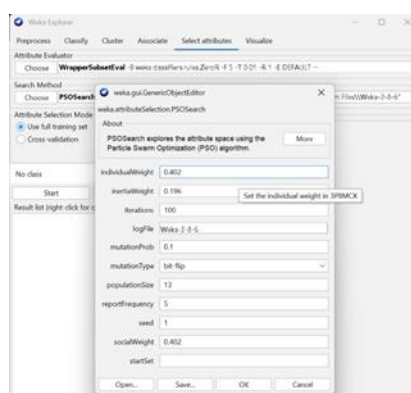


Figura #4: Parámetros Particle Swarm Optimisation (Extraído de Weka)

Con respecto a los MLCs involucrados en el rendimiento de los wrappers, RF, que es parametrizable para iteraciones, se configuró con 50 iteraciones para GA y 100 iteraciones para PSO, esto debido al costo computacional asociado con cada algoritmo metaheurístico. Un cambio implementado para todos los MLCs fue la métrica de evaluación, que se ajustó al AUC, junto con el uso de la validación cruzada de diez iteraciones.

Los parámetros (MLCs) se optimizaron para el mejor subconjunto de características obtenido mediante el método wrapper. Esto se aplicó a todos los modelos, excepto QDA, que no requiere parametrización. Para el modelo kNN, se determinó que el número óptimo de vecinos k , estaba

en el rango impar de 3 a 9. En el caso del modelo MLP, se utilizó una tasa de aprendizaje de 0.3 y un momento de 0.2. Finalmente, para SVM, se empleó un kernel lineal para su configuración.

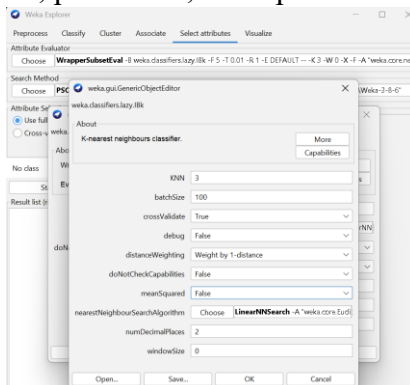


Figura #5: Parámetros de los MLC (Extraído de Weka)

Criterio de selección

Se calculó la media y la desviación estándar de las métricas: área bajo la curva (AUC), precisión (PRE) y sensibilidad (REC). Esto con el fin de evaluar el rendimiento del método propuesto en las fases de entrenamiento y prueba, dado que el conjunto de datos experimental está casi balanceado, la métrica AUC promedio es una medida confiable para evaluar los resultados (Shatnawi, 2024). Por esta razón, se utilizó AUC como medida principal en el Test-U, de esta manera se facilita la elección de un valor pivote, identificando envolturas con diferencias estadísticamente significativas por debajo de un umbral de 0.05. En la clasificación de sismos, se elige el subconjunto wrapper más adecuado para clasificar los mecanismos focales sísmicos.

Software

Para el análisis y clasificación de datos en este trabajo, se utilizó el software Weka. Según Witten (2016), Weka es un software de código abierto que proporciona herramientas para la minería de datos y el aprendizaje automático, incluyendo preprocesamiento de datos, clasificación, regresión, agrupamiento y selección de atributos.

RESULTADOS

En base a la información expuesta en la anterior sección, 42 esquemas de clasificación (wrapper + MLC) fueron evaluados durante la fase de entrenamiento del método propuesto utilizando un método de validación cruzada de diez pliegues. Al mejor método se evaluará usando el 10% restante. Como tal, los resultados obtenidos sobre la base de las métricas computadas se proporcionan en la Tabla 1.

Metaheuristic	Function	L	Subset	MLC	Parameters	AUC±SD	PRE±SD	REC±SD	T-TEST
GA (it=500) pop=13 cl=13 mr=0.2 np=10	NB	12	$f^1, f^2, f^3, f^4, f^8, f^{13}, f^{14}, f^{17}, f^{19}, f^{21}, f^{22}, f^{25}$	QDA	$r=1 \times 10^{-6}$	0.78 ± 0.06	0.65 ± 0.07	0.66 ± 0.07	$p < 0.05$
				kNN	$k=3$	0.78 ± 0.06	0.66 ± 0.07	0.65 ± 0.07	$p < 0.05$
				kNN	$k=5$	0.80 ± 0.05	0.67 ± 0.08	0.66 ± 0.07	$p > 0.05$
				kNN	$k=7$	0.80 ± 0.05	0.68 ± 0.07	0.66 ± 0.06	$p > 0.05$
				kNN	$k=9$	0.81 ± 0.05	0.68 ± 0.08	0.66 ± 0.07	$p > 0.05$
				MLP	$M=0.3 \quad L=0.2$	0.78 ± 0.06	0.64 ± 0.08	0.62 ± 0.07	$p < 0.05$
	RF (it=50)	15	$f^5, f^6, f^7, f^8, f^{11}, f^{13}, f^{15}, f^{16}, f^{17}, f^{18}, f^{20}, f^{21}, f^{22}, f^{23}, f^{25}$	SVM	$KT=linear$	0.69 ± 0.05	0.67 ± 0.07	0.62 ± 0.06	$p < 0.05$
				QDA	$r=1 \times 10^{-6}$	0.78 ± 0.06	0.63 ± 0.08	0.58 ± 0.08	$p < 0.05$
				kNN	$k=3$	0.80 ± 0.05	0.66 ± 0.07	0.65 ± 0.07	$p > 0.05$
				kNN	$k=5$	0.81 ± 0.05	0.69 ± 0.07	0.68 ± 0.06	$p > 0.05$
				kNN	$k=7$	0.82 ± 0.05	0.69 ± 0.06	0.68 ± 0.06	$p > 0.05$
				kNN	$k=9$	0.82 ± 0.05	0.69 ± 0.06	0.67 ± 0.06	$p > 0.05$
	LDA	14	$f^1, f^5, f^9, f^{11}, f^{12}, f^{13}, f^{14}, f^{15}, f^{16}, f^{17}, f^{20}, f^{21}, f^{22}, f^{24}$	MLP	$M=0.3 \quad L=0.2$	0.79 ± 0.05	0.66 ± 0.06	0.65 ± 0.05	$p < 0.05$
				SVM	$KT=linear$	0.69 ± 0.05	0.68 ± 0.08	0.62 ± 0.06	$p < 0.05$
				QDA	$r=1 \times 10^{-6}$	0.80 ± 0.06	0.67 ± 0.07	0.66 ± 0.06	$p > 0.05$
				kNN	$k=3$	0.80 ± 0.05	0.70 ± 0.07	0.68 ± 0.07	$p > 0.05$
				kNN	$k=5$	0.82 ± 0.05	0.71 ± 0.06	0.69 ± 0.06	$p > 0.05$
				kNN	$k=7$	0.82 ± 0.05	0.68 ± 0.06	0.67 ± 0.06	$p > 0.05$
PSO (it=100) pop=13 cl=25 $c_1=0.402$ $c_2=0.402$ $i=0.196$	NB	9	$f^1, f^2, f^4, f^8, f^{13}, f^{14}, f^{19}, f^{22}, f^{25}$	kNN	$k=9$	0.82 ± 0.05	0.67 ± 0.07	0.66 ± 0.06	$p > 0.05$
				MLP	$M=0.3 \quad L=0.2$	0.76 ± 0.05	0.62 ± 0.07	0.61 ± 0.06	$p < 0.05$
				SVM	$KT=linear$	0.66 ± 0.06	0.62 ± 0.09	0.59 ± 0.07	$p < 0.05$
				QDA	$r=1 \times 10^{-6}$	0.76 ± 0.05	0.63 ± 0.07	0.61 ± 0.06	$p < 0.05$
				kNN	$k=3$	0.77 ± 0.06	0.64 ± 0.07	0.63 ± 0.07	$p < 0.05$
				kNN	$k=5$	0.79 ± 0.06	0.67 ± 0.07	0.66 ± 0.07	$p > 0.05$
	RF (it=100)	13	$f^4, f^5, f^6, f^7, f^8, f^{10}, f^{11}, f^{15}, f^{18}, f^{20}, f^{21}, f^{23}, f^{25}$	kNN	$k=7$	0.81 ± 0.05	0.68 ± 0.06	0.67 ± 0.06	$p > 0.05$
				kNN	$k=9$	0.82 ± 0.05	0.70 ± 0.06	0.68 ± 0.05	$p > 0.05$
				MLP	$M=0.3 \quad L=0.2$	0.80 ± 0.05	0.66 ± 0.05	0.65 ± 0.05	$p > 0.05$
				SVM	$KT=linear$	0.70 ± 0.05	0.68 ± 0.09	0.63 ± 0.07	$p < 0.05$
				QDA	$r=1 \times 10^{-6}$	0.79 ± 0.05	0.65 ± 0.07	0.59 ± 0.08	$p > 0.05$
				kNN	$k=3$	0.79 ± 0.05	0.64 ± 0.07	0.63 ± 0.07	$p < 0.05$
	LDA	14	$f^1, f^4, f^5, f^9, f^{10}, f^{11}, f^{12}, f^{13}, f^{14}, f^{17}, f^{21}, f^{22}, f^{24}, f^{25}$	kNN	$k=5$	0.82 ± 0.05	0.70 ± 0.06	0.69 ± 0.06	$p > 0.05$
				kNN	$k=7$	0.83 ± 0.05	0.71 ± 0.06	0.69 ± 0.06	$p > 0.05$
				kNN	$k=9$	$0.83 \pm 0.05^*$	0.69 ± 0.06	0.68 ± 0.06	$p=1$
				MLP	$M=0.3 \quad L=0.2$	0.80 ± 0.05	0.66 ± 0.07	0.65 ± 0.07	$p > 0.05$
				SVM	$KT=linear$	0.71 ± 0.04	0.70 ± 0.07	0.65 ± 0.05	$p < 0.05$
				QDA	$r=1 \times 10^{-6}$	0.79 ± 0.06	0.66 ± 0.07	0.62 ± 0.07	$p > 0.05$

L - subset cardinality; SD - standard deviation; it - number of iterations (epochs); pop - population size; cl - chromosome length
mr - mutation range; np - number of parents; c - acceleration constant; i - inertia; r - ridge; k - number of neighbors; M - momentum
L - learning Rate; KT - kernel type

Tabla #1: Resultados de desempeño del método propuesto

Rendimiento de la fase del método propuesto

El rendimiento de wrappers basados en algoritmos metaheurísticos mostró una reducción efectiva del espacio de características hacia subconjuntos esenciales, mejorando la clasificación del modelo. Los wrappers basados en PSO (NB=9, RF=13, LDA=14) lograron una reducción de atributos más eficiente que los basados en GA (NB=12, RF=15, LDA=14), sin afectar negativamente el desempeño de los Machine Learning Classifiers (MLCs).

El desempeño de los MLCs mostró resultados sólidos, con valores de AUC que oscilaron entre 0.66 y 0.83, confirmando la efectividad de la reducción de características en la clasificación de movimientos sísmicos. Cabe destacar que el clasificador kNN tuvo el mejor rendimiento de forma consistente en todos los wrappers, especialmente con valores pequeños de k , gracias a su capacidad para aprovechar la estructura local de los datos. Por ejemplo, el kNN con $k=5$ en el wrapper GA+LDA alcanzó un AUC de 0.82 ± 0.05 , superando a otros MLCs en la misma configuración y además acercándose al valor que fue escogido como el pivote que tuvo un AUC de 0.83 ± 0.05 (se indica con un asterisco en la Tabla 1). Por el contrario, SVM mostró el menor rendimiento debido a su kernel lineal. Dado que los datos sísmicos a menudo presentan patrones complejos y no lineales, un kernel más flexible podría mejorar los resultados. Esto se evidenció en el wrapper PSO+NB+SVM, donde el AUC fue el más bajo (0.66 ± 0.06), lo que muestra sus limitaciones para manejar la complejidad reducida de los datos. En términos de sensibilidad (REC) y precisión (PRE), kNN también superó a otros clasificadores, reforzando su efectividad con datos sísmicos. Aunque MLP mostró un AUC razonable (0.76-0.79), no superó a kNN, probablemente debido a las necesidades de hiperparámetros y la reducción de características. Más allá de las comparaciones de AUC, las pruebas T confirmaron la significancia estadística de las diferencias entre los MLCs, con valores p inferiores a 0.05 que resaltan los clasificadores que generaron menos falsos positivos. Esto respalda el desempeño superior de kNN. El mejor

resultado destaca su efectividad en la clasificación además de su costo computacional mínimo (ver Tabla 1, resaltado en negrita).

Mejor resultado del método propuesto

Utilizando el método previamente identificado como el mejor PSO+NB+kNN ($k=5$), se realizó una prueba sobre el conjunto completo de datos para evaluar su capacidad de clasificación en los tres tipos de mecanismos focales. Según la matriz de confusión presentada en la Figura 6, se calculó el porcentaje de aciertos presentada en la Tabla 2, en esta se puede observar que el modelo tiene un porcentaje de aciertos promedio del 65%. Este resultado es prometedor, especialmente considerando que el espacio de características original (25 variables) fue reducido en un 64%, dejando únicamente 9 variables relevantes para la clasificación. Al examinar la matriz de confusión, se evidencia que la mayor tasa de aciertos corresponde a la clase "RO", con un total de 119 casos correctamente clasificados. Esto sugiere que el modelo tiene una mayor capacidad para identificar correctamente eventos de esta clase. Esto es posible debido a patrones más marcados o características distintivas en los datos seleccionados que permite una mejor diferenciación sobre las demás. Por otro lado, las clases "SS" y "NO" muestran tasas de acierto más moderadas (111 y 48 casos respectivamente), lo que podría estar relacionado con una mayor superposición en sus atributos o una menor representación de características clave tras la reducción.

a	b	c	<-- classified as
17	0	1	a = SS
3	4	2	b = NO
3	1	14	c = RO

Figura #6: Matriz de confusión PSO+NB+kNN ($k=5$)

SS	NO	RO	% Accuracy	% Accuracy AVG.
17	0	1	94.4	
3	4	2	44.4	72.2
3	1	14	77.8	

SS - Strike Slip; NO - Normal Oblique; RO - Reverse Oblique

Tabla #2: Porcentaje de aciertos de eventos sísmicos

Comparando con la Figura 7, el cual representa al modelo que mejores resultados da, se evidencia que la Figura 6 si bien no presenta la mejor consistencia en comparación con un modelo mucho más estable debido a la cantidad de características, el modelo demuestra una comparación similar en algunas clases. Por ejemplo, la clase SS y RO presenta un número igual en la clasificación de eventos, por otro lado, en la clase NO la diferencia es mucho más notable. Esto confirma que el modelo elegido si bien no es superior, logra un equilibrio entre la reducción de características y la efectividad del modelo, resaltando así su potencial como herramienta robusta para clasificar mecanismos focales en contextos donde la alta reducción de características y el costo computacional sean requisitos para dar una respuesta rápida a eventos sísmicos, como por ejemplo un sistema de alerta temprana.

a	b	c	<-- classified as
17	0	1	a = SS
0	7	2	b = NO
4	0	14	c = RO

Figura #7: Matriz de confusión GA+RF+kNN (k=3)

Relevancia de características

En la Tabla 3, se observa que los wrappers con mejor desempeño en términos de rendimiento AUC son los que utilizan RF en su wrapper, con puntajes de 0.855 y 0.856 en los métodos GA y PSO, respectivamente. Sin embargo, estos modelos utilizan 15 y 13 variables, lo que demuestra que una reducción de características superior al 50% podría afectar ligeramente la efectividad en comparación con otros clasificadores. Esta menor reducción de características (menos del 50% del total) podría atribuirse a varias propiedades inherentes de RF. Sin embargo, se puede notar que el modelo seleccionado como el mejor, PSO+NB, muestra un desempeño consistente con la tendencia de los demás, lo cual se explica por su alta capacidad de reducción en el espacio de características.

Wrapper		Merit (Mean AUC)
Searcher	Classifier	
GA	NB	0,782
	RF	0,855
	LDA	0,781
PSO	NB	0,775
	RF	0,856
	LDA	0,777

Tabla #3: Evaluación de rendimiento

Además, se observó una tendencia interesante en todos los resultados y es que en dichos modelos aparece un conjunto de variables recurrentes, específicamente en cuatro de los seis modelos evaluados los cuales son: máximo (f_5), oblicuidad (f_8), centro de masa en x (f_{11}), centroide x (f_{13}), circularidad (f_{17}), elongación (f_{18}), entropía (f_{21}), contraste (f_{22}) y momento inverso (f_{25}). Estas características representan atributos comunes seleccionados desde diferentes perspectivas, sugiriendo su alta relevancia para la clasificación de eventos sísmicos. Basado en investigaciones previas, se observa que la mayoría de los atributos recurrentes describen la forma del terremoto en 2D. Esto sugiere que las características de descriptores de forma determinados por los eventos sísmicos, como el centroide o el centro de masa provocan un comportamiento distintivo en los gráficos 2D según el tipo de mecanismo focal que se analiza. Ver Figuras 8, 9 y 10.

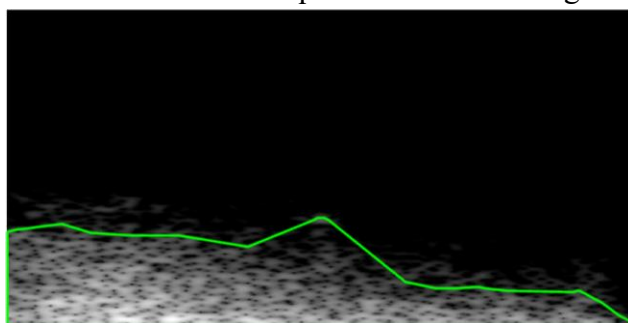


Figura #8: Forma 2D del mecanismo focal SS

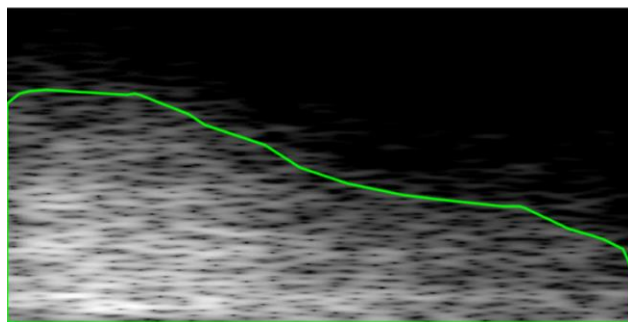


Figura #9: Forma 2D del mecanismo focal NO

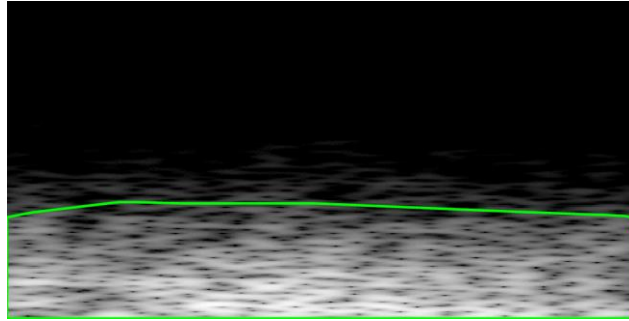


Figura #10: Forma 2D del mecanismo focal RO

CONCLUSIONES

Este estudio demuestra como la efectividad de un enfoque híbrido donde se combina algoritmos metaheurísticos con clasificadores de aprendizaje automático para la reducción de características y la clasificación de señales sísmicas. Los resultados destacan que PSO es más eficiente que GA para seleccionar características relevantes, y que kNN, con valores bajos de k , es el clasificador más adecuado para datos locales.

Este análisis representa un avance significativo no solo para el campo de machine learning y su utilidad en diversas disciplinas, sino también en el análisis y entendimiento del movimiento sísmico, lo que es de vital importancia en la ingeniería civil. Por ejemplo, comprender cómo responden diferentes tipos de estructuras ante ciertos patrones de movimiento sísmico podría guiar el desarrollo de diseños optimizados para edificios, puentes y otras infraestructuras críticas. En regiones propensas a terremotos como Ecuador, este tipo de análisis puede ser crucial para el diseño de estructuras más resistentes y la implementación de sistemas de monitoreo en tiempo real. Esto convierte al machine learning en una herramienta clave para abordar los desafíos que plantea la sismología en la ingeniería civil moderna.

Más allá de los resultados obtenidos, este trabajo plantea preguntas importantes sobre cómo futuras aplicaciones pueden transformar la forma en que abordamos los fenómenos naturales y diseñamos soluciones resilientes. Con una evolución constante en los métodos y bases de datos, el horizonte de investigación y aplicación sigue expandiéndose, dejando abierta la posibilidad de descubrir nuevas formas de proteger y fortalecer las comunidades frente a los retos que plantea la naturaleza.

REFERENCIAS BIBLIOGRÁFICAS

- Aki, K., & Richards, P. G. (2002). Quantitative seismology (2nd ed.). University Science Books.
- Allen, R. M., & Melgar, D. (2019). Earthquake early warning: Advances, scientific challenges, and societal needs. *Annual Review of Earth and Planetary Sciences*, 47(1), 361–388. <https://doi.org/10.1146/annurev-earth-053018-060232>
- Ancheta, T. D., Darragh, R. B., Stewart, J. P., Seyhan, E., Silva, W. J., Chiou, B. S., ... & Boore, D. M. (2013). PEER NGA-West2 database. Pacific Earthquake Engineering Research Center, Berkeley, CA.
- Bayliss, K., Naylor, M., Illian, J., & Main, I. G. (2020). Data-driven optimization of seismicity models using diverse data sets: Generation, evaluation, and ranking using inlabru. *Journal of Geophysical Research: Solid Earth*, 125(11), e2020JB020226.
- Bayliss, K., Naylor, M., Illian, J., & Main, I. G. (2020). Data-driven optimization of seismicity models using diverse data sets: Generation, evaluation, and ranking using inlabru. *Journal of Geophysical Research: Solid Earth*, 125(11), e2020JB020226. <https://doi.org/10.1029/2020JB020226>
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87. <https://doi.org/10.1145/2347736.2347755>
- Fontanari, T., Fróes, T. C., & Recamonde-Mendoza, M. (2022). Cross-validation strategies for balanced and imbalanced datasets. In *Brazilian Conference on Intelligent Systems* (pp. 626–640). Springer. https://doi.org/10.1007/978-3-030-89654-8_43
- Ghamisi, P., & Benediktsson, J. A. (2014). Feature selection based on hybridization of genetic algorithm and particle swarm optimization. *IEEE Geoscience and Remote Sensing Letters*, 12(2), 309–313. <https://doi.org/10.1109/LGRS.2014.2350537>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(Mar), 1157–1182.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.
- Kupervasser, O. (2014). The mysterious optimality of naive bayes: Estimation of the probability in the system of “classifiers.” *Pattern Recognition and Image Analysis*, 24, 1–10. <https://doi.org/10.1134/S1054661814010030>
- Li, S., Zhang, K., Chen, Q., Wang, S., & Zhang, S. (2020). Feature selection for high-dimensional data using weighted k-nearest neighbors and genetic algorithm. *IEEE Access*, 8, 139512–139528. <https://doi.org/10.1109/ACCESS.2020.3013793>

- Lomnitz, C. (2013). *Global tectonics and earthquake risk*. Elsevier.
- McCall, J. (2005). Genetic algorithms for modelling and optimisation. *Journal of Computational and Applied Mathematics*, 184(1), 205–222. <https://doi.org/10.1016/j.cam.2004.08.028>
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Mousavi, S. M., Sheng, Y., Zhu, W., & Beroza, G. C. (2019). Stanford earthquake dataset (STEAD): A global data set of seismic signals for AI. *IEEE Access*, 7, 179464–179476. <https://doi.org/10.1109/ACCESS.2019.2956311>
- Naveen, P., & Diwan, B. (2020). Relative analysis of ML algorithm QDA, LR, and SVM for credit card fraud detection dataset. En *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 976–981). IEEE. <https://doi.org/10.1109/I-SMAC48924.2020.9240377>
- Salam, M. A., Ibrahim, L., & Abdelminaam, D. S. (2021). Earthquake prediction using hybrid machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 12(5), 654–652. <https://doi.org/10.14569/IJACSA.2021.0120567>
- Scholz, C. H. (2019). *The mechanics of earthquakes and faulting*. Cambridge University Press.
- Shatnawi, R. (2024). Deriving change-prone thresholds from software evolution using ROC curves. *The Journal of Supercomputing*, 80(16), 23565–23591. <https://doi.org/10.1007/s11227-024-04748-2>
- Theurer, M. M., Jiménez, J. V., Velasco, G. M., & Zambrano, I. B. (2017). Análisis sobre la recurrencia de terremotos severos en Ecuador. *Prisma tecnológico*, 8(1), 12–17. <https://doi.org/10.33262/pt.v8i1.244>
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.
- Xanthopoulos, P., Pardalos, P. M., Trafalis, T. B., Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. In *Robust Data Mining* (pp. 27–33).
- Xue, B., Zhang, M., & Browne, W. N. (2012). Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE Transactions on Cybernetics*, 43(6), 1656–1671. <https://doi.org/10.1109/TCYB.2012.2225042>
- Yépez, F., Benítez, D., Pérez, N., & Grijalva, F. (2024). Earthquake fault mechanism prediction using AI algorithms applied to strong motion records. *Proceedings of the 18th World Conference on Earthquake Engineering (WCEE2024)*. Milan, Italy: International Association for Earthquake Engineering (IAEE).
- Zwane, S., Tarwireyi, P., & Adigun, M. (2018). Performance analysis of machine learning classifiers for intrusion detection. En *2018 International Conference on Intelligent and*

Innovative Computing Applications (ICONIC) (pp. 1–5). IEEE.
<https://doi.org/10.1109/ICONIC.2018.8597301>