Clasificación de Eventos Sísmicos: Detección de Terremotos y Explosiones mediante Machine Learning

Constanza Bustos y Juan Sebastián Gutiérrez

1. Introducción

Un terremoto consiste en la liberación repentina de energía acumulada en la corteza terrestre, generando ondas que se propagan en todas direcciones.

El objetivo principal de este estudio es identificar patrones geográficos y físicos que permitan distinguir estos eventos naturales de otros artificiales (como explosiones nucleares o derrumbes). Para ello, se busca construir un modelo predictivo robusto capaz de clasificar eventos "Naturales"vs "No Naturales", superando el desafío que presenta el fuerte desbalance de clases en los datos sismológicos reales.

2. Datos y Preprocesamiento

El conjunto de datos utilizado corresponde a registros del USGS, conteniendo inicialmente 23,412 filas y 21 columnas. Se realizó un análisis exhaustivo para identificar la calidad de la información.

Como se detalla en la Tabla 1, se detectaron variables con una gran cantidad de datos vacíos. Para asegurar la calidad del modelo, se descartaron las columnas marcadas en **azul** (ej. errores de medición horizontal/vertical y datos de estaciones específicas). Además, se eliminaron identificadores administrativos irrelevantes para el análisis físico (ID, Source, Status).

Tabla 1: Conteo de valores nulos. En azul, variables descartadas.

Variable	Nulos	Variable	Nulos
Date	0	Mag Seismic Stations	20848
Time	0	Azimuthal Gap	16113
Latitude	0	Horizontal Distance	21808
Longitude	0	Horizontal Error	22256
Type	0	Root Mean Square	6060
Depth	0	ID	0
Depth Error	18951	Source	0
Depth Seismic Stations	16315	Location Source	0
Magnitude	0	Magnitude Source	0
Magnitude Type	3	Status	0
Magnitude Error	23085		

3. Metodología

3.1. Manejo del Desbalance (SMOTE)

El dataset presentaba un desbalance crítico: más de 23,000 sismos frente a solo 180 eventos no sísmicos. Entrenar un modelo con esta distribución sesgaría los resultados hacia la clase mayoritaria.

Para solucionar esto, aplicamos la técnica SMOTE (*Synthetic Minority Over-sampling Technique*) exclusivamente en el conjunto de entrenamiento. Como muestra la Tabla 2, esto permitió igualar la cantidad de muestras de la clase minoritaria ("No-Earthquake") a la mayoritaria, generando datos sintéticos para un aprendizaje equilibrado.

Tabla 2: Distribución de clases antes y después de SMOTE.

Clasificación	Total Orig.	Train (Pre)	Train (Post)
0: Earthquake	23,229	16,260	16,260
1: No-Earthquake	180	126	16,260

3.2. Configuración del Modelo

Se utilizó un algoritmo de **Random Forest**. Mediante una búsqueda de hiperparámetros, se encontraron los valores óptimos: 100 estimadores (`n_estimators`), un mínimo de 5 muestras para dividir un nodo (`min_samples_split`) y sin límite de profundidad preestablecido (`max_depth=None`).

4. Resultados y Análisis

4.1. Evaluación del Modelo

El modelo optimizado logró resultados notables en el conjunto de prueba. Fue capaz de identificar eventos no sísmicos con un **98 % de recall** y **96 % de precision**, lo cual es excelente considerando que originalmente representaban solo el 0.7 % de los datos.

La Tabla 3 detalla el desempeño por clase, evidenciando un F1-Score casi perfecto.

Tabla 3: Métricas de Clasificación en el Set de Prueba.

Clase	Precision	Recall	F1-score	Support
Earthquake	1.00	1.00	1.00	6969
No-Earthquake	0.96	0.98	0.97	54
Accuracy			**1.00**	7023

La Matriz de Confusión (Fig. 1) confirma la precisión del clasificador: de los 54 eventos no naturales reales en el test, el modelo detectó correctamente 53, cometiendo únicamente un error de falso negativo.

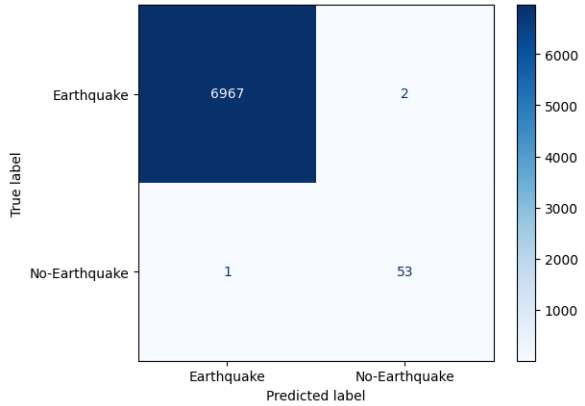

Figura 1: Matriz de Confusión normalizada.

4.2. Importancia de Variables

El análisis de importancia de variables (Fig. 2) revela la lógica física aprendida por el modelo:

- **Dominancia de la Profundidad:** *Depth* es el predictor más influyente. Confirma la hipótesis física: la restricción de profundidad es la "firma"principal de una explosión superficial frente a sismos profundos.

- **Ubicación:** La geolocalización permite asociar eventos a sitios conocidos (ej. polígonos de prueba nucleares).

Finalmente, la Figura 3 ilustra la relación entre el tipo de magnitud y la profundidad, validando que existen patrones distinguibles que el modelo pudo aprovechar para realizar la clasificación.

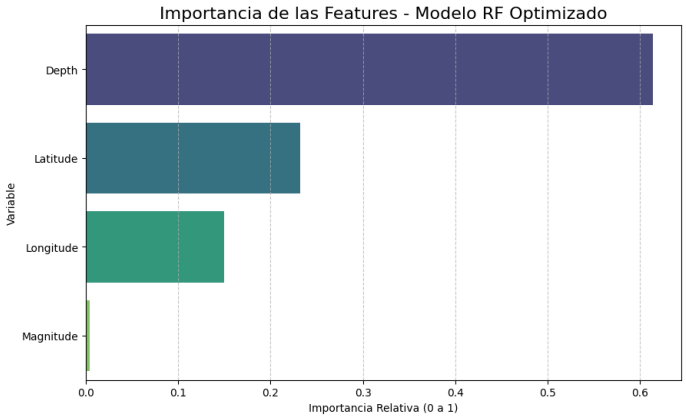

Figura 2: Importancia relativa de las variables en el modelo.

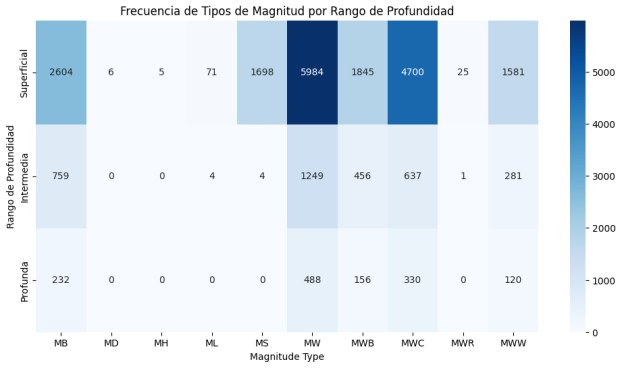

Figura 3: Relación entre Tipo de Magnitud y Profundidad.

5. Conclusiones

1. **Eficacia ante el Desbalance:** La combinación de Random Forest con SMOTE permitió clasificar eventos que representan menos del 1 % de la data con un F1-Score del 97 %.

2. **Validación Física:** El modelo no es una çaja negra"; su dependencia prioritaria en la Profundidad y Ubicación se alinea con los principios geofísicos.

3. **Limitaciones:** El dataset se restringe a magnitudes $> 5{,}5$, limitando el modelo a eventos de energía media-alta. Para microsismicidad, se requeriría análisis de formas de onda completa.