Prediciendo la magnitud de terremotos a lo largo de la historia

Justine Haefele[1], Javiera Martínez[1], Joseph Dabre[1]

[1]Departamento de Física, Universidad Técnica Federico Santa María

1 Introducción

Gran parte del tiempo, la detección de sismos no es posible hasta que ya está sucediendo. Aunque existen sistemas de alerta temprana en zonas de alta actividad, aún no es posible anticipar la magnitud de un sismo antes de que ocurra. Es por este motivo que analizamos datos proporcionados por el Centro Nacional de Información Sísmica de los Estados Unidos (NEIC), con el fin de predecir la magnitud de futuros eventos sísmicos a partir de coordenadas geográficas como la latitud, longitud y profundidad. Para ello, implementamos y comparamos algoritmos de regresión, como Random Forest y K-Nearest Neighbors (kNN), evaluando su capacidad para capturar patrones en los datos y generar predicciones acordes al conocimiento geológico actual.

2 Datos y EDA

El dataset proporcionado por el NEIC incluye un registro de la fecha, hora, coordenadas geográficas, profundidad, magnitud y el origen de todos los eventos sísmicos del mundo con una magnitud igual o superior a 5.5 registrados desde 1965 hasta 2016. Consta de 21 columnas, con un total de 23412 datos.
Al revisar la distribución de cada columna vimos que los eventos sísmicos se clasifican en terremotos, explosiones, explosiones nucleares y estallidos de roca. Ocurren mayoritariamente cerca de la latitud 0°, es decir cercano al Ecuador terrestre, y levemente hacia el sur. También, tienen mayor frecuencia en las longitudes extremas $(150°, -150°)$ y ocurren mayoritariamente a bajas profundidades (< 100 km), donde predominan los sismos de baja intensidad (~ 5.5). Pudimos apreciar que la zona de epicentro del sismo está relacionada con la profundidad, siendo mayor cerca de las placas tectónicas. Los datos cuentan con varios tipos magnitudes, siendo la magnitud de momento (M_W) y sus derivadas las que presentan la mayor cantidad de datos. Para mantener la coherencia en la escala, utilizamos únicamente estos tipos de magnitud. El pre-procesamiento de los datos arrojó que no existen outliers ni datos duplicados que deban ser tratados. Sin embargo, la columna `Magnitude Type` presentó 3 datos NaN, que decidimos eliminar ya que no tenemos forma de saber en qué escala se midieron. De esta manera, trabajamos con un dataframe filtrado de acuerdo a la magnitud de momento M_W.

3 Random Forest Regressor

Random forest es un algoritmo robusto que puede captar interacciones o relaciones no lineales aún cuando la correlación lineal de variables es baja. El dataset es de un tamaño razonable y particularmente no vemos mucho ruido o presencia de outliers, por lo que el sobreajuste, común de los modelos de árboles, está relativamente fuera. Corrimos el modelo de dos formas, las cuales se exponen en la Tabla (1).

Parámetro	**Configuración 1**	**Configuración 2**
n_estimators	100	100
random_state	42	42
max_depth	–	10
min_samples_leaf	–	2
min_samples_split	–	5

Table 1: Parámetros utilizados en las dos configuraciones del modelo Random Forest Regressor.

Los hiper-parámetros fueron optimizados mediante un gridsearch pequeño, intentando encontrar la profundidad, mínimo de muestras antes de dividir y cantidad de muestras por hoja óptimas respecto al problema.

Se utilizaron tres métricas para evaluar el modelo. MAE, que indica cuánto se equivoca el modelo, en promedio al predecir la magnitud. MSE y RMSE que cuantifican el error cuadrático, es decir, miden qué tan lejos están las predicciones de los valores reales.

4 K-nearest neighbors (kNN)

Partimos asumiendo que los sismos que ocurren en condiciones similares tienden a presentar magnitudes parecidas. Por ejemplo, si dos eventos se registran en zonas geográficas cercanas o a profundidades similares, es razonable esperar que sus magnitudes tengan valores cercanos. Bajo esta idea se eligió el modelo KNN, ya que nos permite estimar la magnitud de un sismo basándose en los eventos más cercanos en términos de latitud, longitud y profundidad.

Se realizó un ajuste de hiperparámetros utilizando GridSearchCV para mejorar el desempeño. El mejor modelo se obtuvo con 11 vecinos, distancia Manhattan (p = 1) y pesos uniformes.

Con un MAE promedio de 0.316, MSE y RMSE promedio de 0.181 y 0.425, respectivamente. Finalmente, el coeficiente de determinación refleja qué tan bien el modelo explica la variabilidad de los datos, y se obtuvo un valor promedio de -0.034.

De los valores de las métricas notamos que KNN optimizado tiene errores aceptables en promedio (MAE y RMSE), pero el coeficiente de determinación es negativo lo que significa que el modelo predice con cierta precisión local, pero no generaliza bien. Puede usarse como aproximación, pero habría que considerar modelos más complejos como redes neuronales, o modelos probabilísticos si se busca mejorar la precisión.

5 Discusión y conclusiones

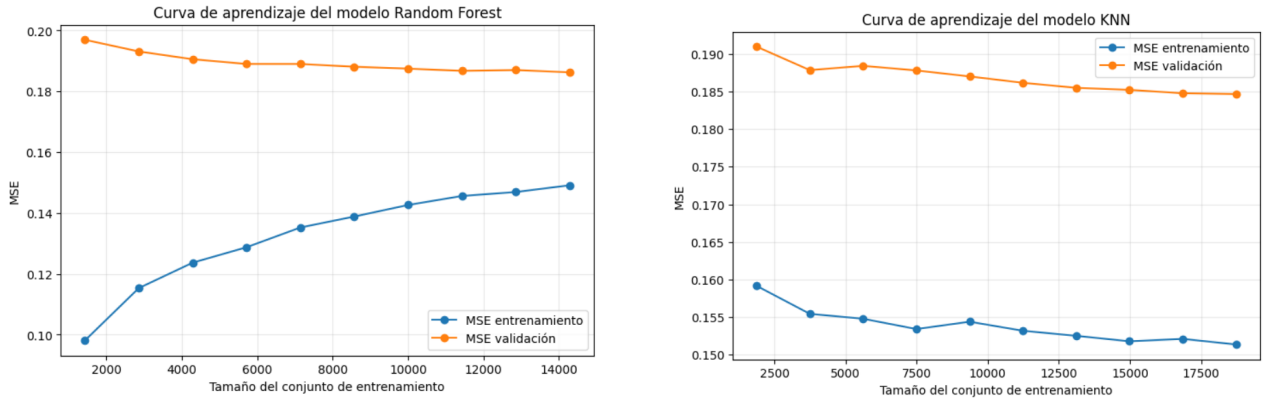

Figure 1: Curvas de aprendizaje de los modelos Random Forest y KNN

En la Figura 5 se presentan las curvas de aprendizaje para los dos modelos. Notamos que para el caso de Random Forest, se observa una brecha entre el error de entrenamiento y el de validación, la cual se reduce gradualmente, indicando que el modelo mejora su capacidad de generalización con más datos. Por otro lado, el modelo KNN presenta una brecha más pequeña, pero al aumentar los datos no hay una mejora. Así Random Forest muestra un mayor potencial de mejora con más datos, mientras que KNN parece alcanzar rápidamente su límite de capacidad predictiva.

Finalmente, aunque Random Forest se comporta mejor que KNN, ambos tienen desempeños insuficientes lo que se debe principalmente a que la magnitud de un sismo no depende directamente de la latitud, la longitud o la profundidad. La magnitud está determinada por procesos físicos internos de ruptura en la corteza terrestre, los cuales son no lineales y en gran medida impredecibles. Incluso dos sismos que suceden muy cerca o a la misma profundidad pueden tener magnitudes completamente distintas, ya que la energía liberada depende de factores como el tamaño de la ruptura, las tensiones acumuladas, el tipo de falla y las propiedades geológicas locales, elementos que no están contenidos en el dataset.