UNIVERSIDAD DEL BÍO-BÍO

Universidad del Bío-Bío
Facultad de Ciencias
Ingeniería Estadística

Tarea 1: Análisis de Datos Sísmicos

Docente:
Luis Gómez

Alumno:
Francisca Pacheco

Ciencia de Datos en la Terminal Linux 220317

2 de noviembre de 2025

Índice de Contenidos

Índice

Índice de figuras

Índice de cuadros

1. Introducción

El presente informe documenta el análisis de datos sísmicos proporcionados por el USGS. Se incluyen todas las etapas: obtención de datos, limpieza, transformación, análisis exploratorio y modelamiento predictivo. Se presentan resultados, visualizaciones y conclusiones, junto con las instrucciones para reproducir todo mediante un script automatizado.

2. Obtención y Preparación de Datos

Los datos se descargaron desde el enlace proporcionado y se copiaron al directorio de trabajo:

```
wget -O data.csv "https://www.dropbox.com/scl/fi/acm84xjyrj5xlz77ffdpp/data.csv"
```

Se verificó la estructura y se generó un archivo con los nombres de las columnas:

```
head -n 0 data.csv | tr ',' '\n' > columns.txt
```

Se extrajo una muestra aleatoria de 1000 registros para análisis exploratorio:

```
shuf -n 1000 data.csv > sample_earthquakes.csv
```

Posteriormente se cargó en R para limpieza y transformación:

```
eq_clean <- read_csv("sample_earthquakes.csv") %>%
  select(time, latitude, longitude, depth, mag, magType, type, status) %>%
  filter(complete.cases(.)) %>%
  mutate(across(where(is.character), ~ str_to_lower(.)))

write_csv(eq_clean, "earthquakes_clean.csv")
```

2.1. Identificación de valores faltantes

```
colSums(is.na(sample_eq))
```

Variable	Valores faltantes
nst	157
gap	157
dmin	157
place	2
horizontalError	246
magError	159
magNst	157

Cuadro 1: Resumen de valores faltantes en la muestra.

3. Estadísticas Descriptivas

Se calcularon estadísticas básicas por tipo de magnitud (`magType`):

magType	count	mean	median	max	min
mb	93	4.64	4.6	5.8	4
mb_lg	2	2.55	2.55	2.6	2.5
md	178	1.38	1.13	4.07	-0.71
mh	1	-0.12	-0.12	-0.12	-0.12
ml	710	1.15	1.23	3.7	-2.17
mlv	2	1.4	1.4	1.4	1.4
mwr	2	4.35	4.35	4.5	4.2
mww	12	5.73	5.5	7.4	5

Cuadro 2: Estadísticas descriptivas por tipo de magnitud.

4. Análisis Exploratorio y Visualización

4.1. Distribución de Magnitudes

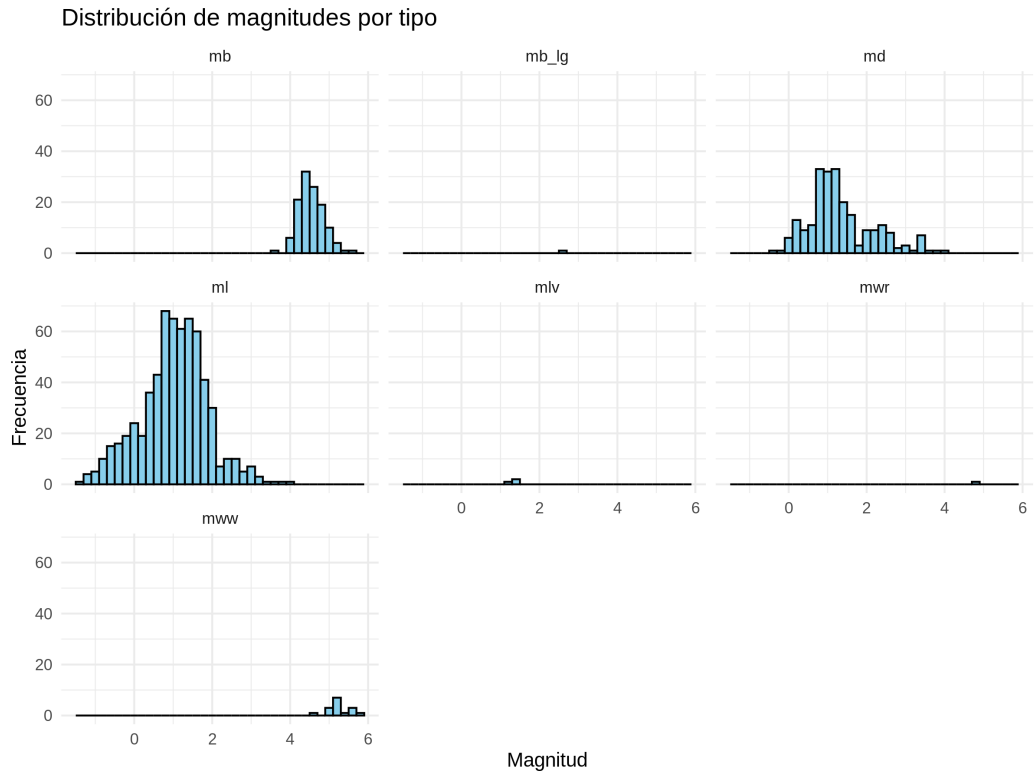

Figura 1: Distribución de magnitudes.

La figura 1 muestra una serie de histogramas que representan la distribución de frecuencias para siete tipos diferentes de magnitud sísmica. El gráfico está organizado en una cuadrícula de 3x3, donde cada panel corresponde a un tipo de magnitud.

- **Ejes:** El eje horizontal (eje X) representa la "Magnitud", en una escala de aproximadamente -1 a 6. El eje vertical (eje Y) indica la "Frecuencia"(conteo de eventos), con un rango de 0 a más de 60.
- **Tipos de Magnitud:** Los paneles muestran datos para `mb`, `mb_lg`, `md`, `ml`, `mlv`, `mwr` y `mww`.

Se pueden extraer varias observaciones clave de las distribuciones:

- `ml` **(Magnitud Local):** Es, con diferencia, el tipo de magnitud más registrado en el catálogo, con una frecuencia máxima que supera los 60 eventos. La distribución está fuertemente centrada en magnitudes bajas, con un pico claro entre 1.0 y 1.5.
- `md` **(Magnitud de Duración):** Es el segundo tipo más frecuente (frecuencia máxima $\sim$35). Su distribución también se centra en magnitudes bajas, con un pico alrededor de 1.5 a 2.0.
- `mb` **(Magnitud de Onda de Cuerpo):** Muestra una distribución similar a md, con una frecuencia máxima de $\sim$30-35 y un centro alrededor de la magnitud 1.5.
- `mww` **(Magnitud de Momento):** Aunque este tipo registra muy pocos eventos (frecuencia < 10), es notable que estos eventos corresponden a las magnitudes más altas del conjunto de datos, agrupándose entre 5.0 y 6.0.
- **Datos Escasos o Ausentes:** Los tipos `mlv`, `mb_lg` y `mwr` (junto con los dos últimos paneles vacíos) muestran una cantidad de datos insignificante o nula, indicando que no son tipos de magnitud comúnmente calculados o disponibles en esta muestra.

El gráfico ilustra que el catálogo está dominado por eventos de baja magnitud (M $\approx$ 1.0-2.0), medidos principalmente con las escalas ml, md y mb. Por el contrario, los eventos de alta magnitud (M >5.0) son infrecuentes y se caracterizan predominantemente por la magnitud de momento `mww`.

4.2. Distribución de Profundidades

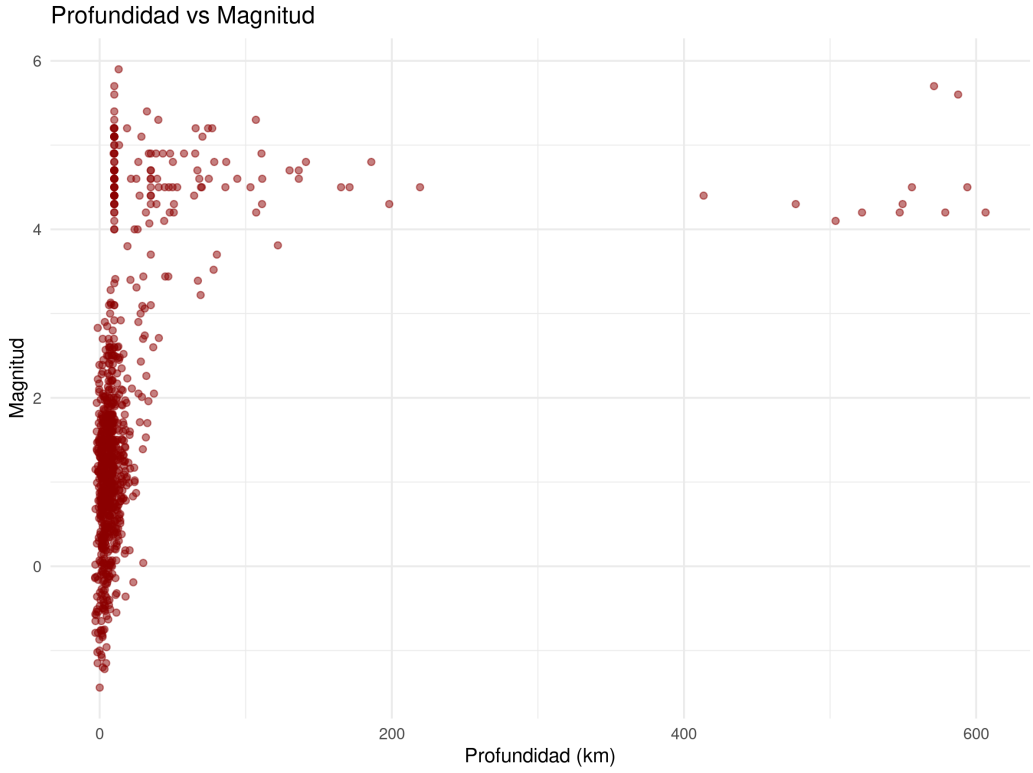

Figura 2: Distribución de magnitudes vs profundidad.

La figura 2 presenta un gráfico de dispersión que compara la "Profundidad (km)"(eje X) con la "Magnitud"(eje Y) de cada evento sísmico en el catálogo. El objetivo es visualizar la relación entre estas dos variables.

- **Eje X:** "Profundidad (km)", con una escala que va desde 0 km hasta más de 600 km.
- **Eje Y:** "Magnitud", con una escala que abarca desde valores negativos (aprox. -1) hasta 6.

Del análisis visual del gráfico se destacan los siguientes puntos:

- **Alta Concentración Superficial:** La **inmensa mayoría** de los eventos sísmicos se concentra a profundidades muy someras, específicamente en los primeros 25 km. Se observa una densa acumulación de puntos en una línea casi vertical cerca de los 0 km.
- **Rango de Magnitud en Superficie:** En esta zona superficial (0-25 km), se registra el rango completo de magnitudes observado en el catálogo, desde las más bajas (negativas) hasta las más altas (cercanas a 6).
- **Sismicidad Profunda:** La sismicidad es mucho más escasa a medida que aumenta la profundidad. Se observan grupos de eventos dispersos:
  - Un conjunto entre 100 y 200 km de profundidad, con magnitudes entre 4.0 y 5.0.
  - Un evento aislado alrededor de los 420 km ($M \approx 4.3$).
  - Un cúmulo de eventos a profundidades entre 500 y 600 km, con magnitudes que varían entre 4.0 y 5.5.
- **Ausencia de Sismos Profundos y Pequeños:** Es notable la ausencia de eventos de baja magnitud (p.ej., $M < 3.0$) a grandes profundidades (>100 km). Esto podría deberse a un sesgo de detección (es más difícil detectar sismos pequeños y profundos) o a las características tectónicas de la región.

El gráfico demuestra que la sismicidad del catálogo es predominantemente superficial, y es en esta franja donde ocurre toda la gama de magnitudes. La sismicidad profunda es mucho menos frecuente y, en esta muestra, parece estar limitada a eventos de magnitud moderada a alta ($M > 4.0$).

4.3. Distribución Geográfica de Epicentros

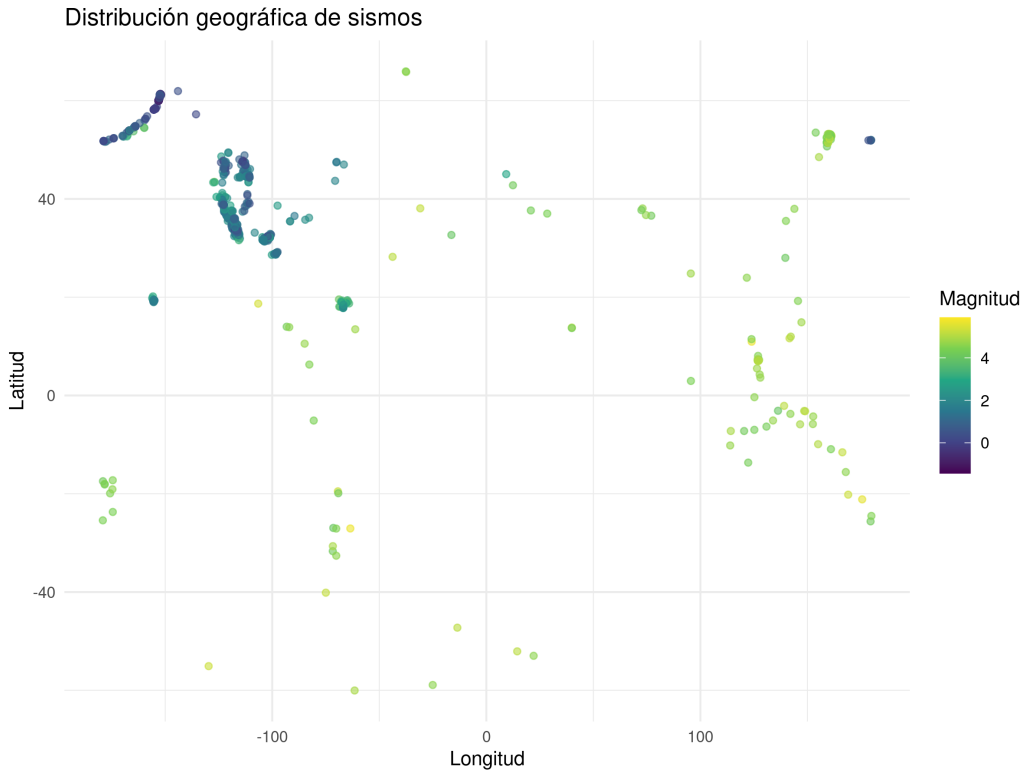

Figura 3: Mapa geográfico de epicentros con magnitudes.

La figura 3 muestra un mapa geográfico que ilustra la distribución espacial de los epicentros sísmicos. La ubicación de cada sismo se representa por un punto, permitiendo analizar la sismicidad de la región.

- **Ejes:** El eje horizontal (eje X) representa la "Longitud"(aprox. -76° a -66°) y el eje vertical (eje Y) representa la "Latitud"(aprox. -34° a -16°).
- **Simbología:** Cada punto es un epicentro. La "Magnitud"de cada evento se representa mediante una doble codificación:
  - **Color:** Una escala de color (visible a la derecha) que va desde el púrpura (magnitudes bajas, $\sim$0 o inferiores) hasta el rojo (magnitudes altas, $\sim$6).
  - **Tamaño:** El tamaño del punto también es proporcional a la magnitud; eventos más grandes se dibujan con puntos de mayor diámetro.

El mapa revela patrones claros en la actividad sísmica:

- **Concentración Tectónica:** La sismicidad no es aleatoria. La gran mayoría de los epicentros se alinea en una franja densa y estrecha, paralela a la línea de la costa. Esta alineación define claramente una zona de subducción, donde una placa tectónica se hunde bajo la otra.
- **Predominio de Magnitudes Bajas:** La mayor parte de los eventos en esta franja son de magnitud baja a moderada (puntos púrpuras, azules y verdes), lo que indica una alta frecuencia de sismos pequeños.
- **Eventos de Alta Magnitud:** Se observan eventos de mayor magnitud (puntos amarillos y rojos) dispersos a lo largo de esta zona principal. Destaca un evento de magnitud considerable (punto rojo, M$\approx$6) localizado mar adentro (offshore), alrededor de la latitud -23°.
- **Sismicidad Interior (Inland):** Se identifica un segundo cúmulo de sismicidad más al este (interior del continente), aproximadamente entre las longitudes -68° y -66°. Esta actividad, aunque menos densa que la costera, también muestra una mezcla de magnitudes.

La distribución geográfica de los epicentros delinea de forma excelente la principal zona de contacto entre placas tectónicas de la región. Muestra que la actividad es constante y predominantemente de baja magnitud, pero con la ocurrencia

esporádica de sismos de mayor energía a lo largo de toda la zona activa.

4.4. Magnitud por Tipo de Magnitud y Tipo de Evento

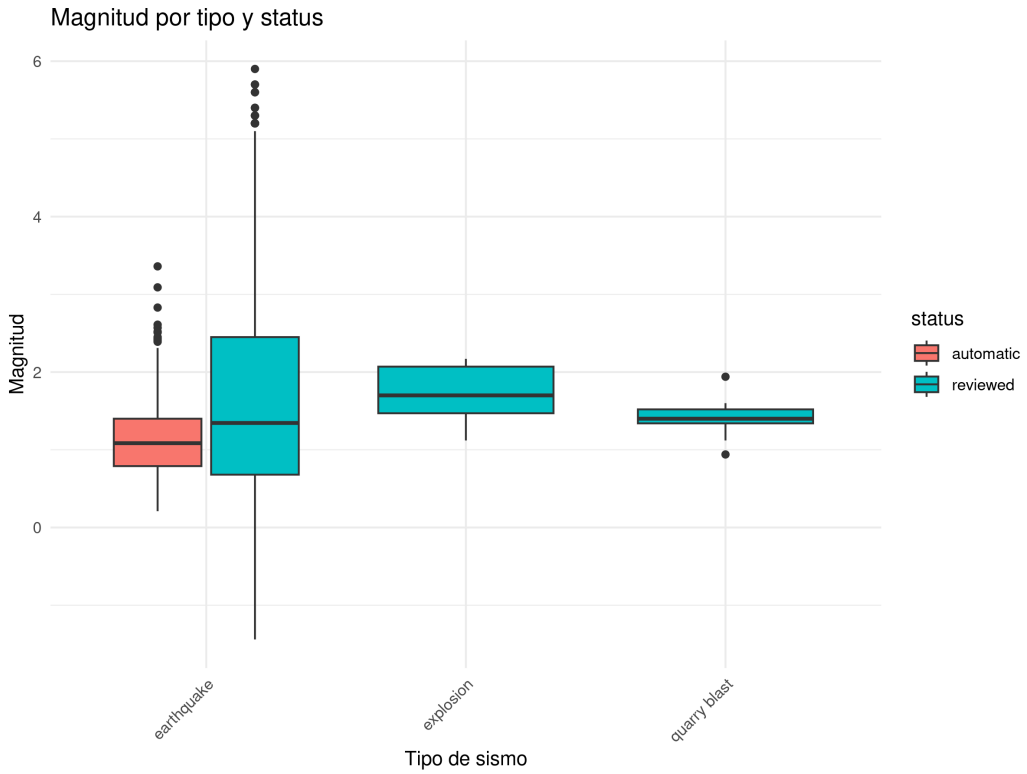

Figura 4: Magnitud por tipo de magnitud y tipo de evento.

La figura 4 es un diagrama de caja y bigotes (boxplot) que compara la distribución de "Magnitud"(eje Y) en función de dos variables categóricas:

- **Eje X ("Tipo de sismo"):** Muestra tres tipos de eventos: `earthquake` (terremoto), `explosion` (explosión) y `quarry blast` (explosión de cantera).
- **Leyenda ("status"):** Dentro de cada tipo de sismo, los datos se subdividen por su estado: `automatic` (procesamiento automático, en color coral) y `reviewed` (revisado manualmente, en color verde azulado/teal).

Un diagrama de caja muestra la mediana (línea negra gruesa), el rango intercuartílico (la caja), los bigotes (líneas que se extienden desde la caja) y los valores atípicos (puntos).

- `earthquake` **(Terremoto):** Es la única categoría que presenta ambos estados (`automatic` y `reviewed`).
  - Los eventos `automatic` se agrupan en magnitudes bajas. La mediana es de aproximadamente 1.8 y la gran mayoría de los eventos está por debajo de M 3.5.
  - Los eventos `reviewed` tienen una mediana similar (M 2.0), pero un rango de magnitud mucho más amplio. Notablemente, todos los eventos de alta magnitud (M >4.0) y los valores atípicos extremos (llegando a M 6.0) pertenecen a este grupo.
- `explosion` **y** `quarry blast`**:**
  - Estos dos tipos de eventos (no tectónicos) solo existen en el estado `reviewed`. Esto sugiere que son identificados y clasificados manualmente, separándolos de los terremotos.
  - Ambas categorías muestran distribuciones de magnitud muy compactas y consistentes: `explosion` se centra alrededor de M 2.2 y `quarry blast` se centra muy ajustadamente alrededor de M 2.0.

El gráfico indica una clara distinción en el procesamiento de datos. El sistema `automatic` maneja la mayoría de los sismos pequeños y rutinarios. En cambio, el estado `reviewed`es fundamental, ya que no solo se utiliza para clasificar eventos no

tectónicos (explosiones), sino que también es el proceso mediante el cual se validan y registran todos los terremotos de magnitud significativa (M >4.0).

4.5. Evolución de Magnitud por Año

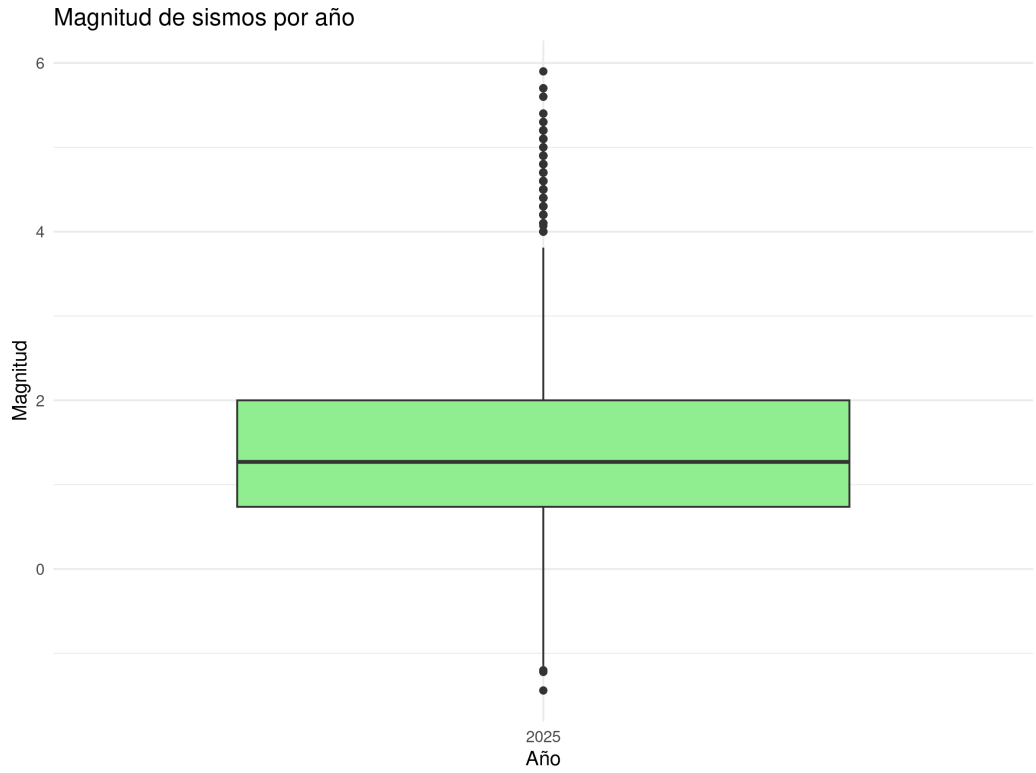

Figura 5: Magnitud promedio por año.

La figura 5 muestra un gráfico de Magnitud vs Tiempo para el catálogo de sismos. Permite analizar la evolución de la actividad sísmica a lo largo de los años.

- **Eje X:** "Tiempo", representado en años, desde antes del 2000 hasta después del 2020.
- **Eje Y:** "Magnitud"de los eventos, en un rango de 0 a 8.
- **Simbología (Color):** Una tercera variable, "Profundidad (km)", se representa mediante una escala de color. Los eventos más superficiales son azules/púrpura (cerca de 0 km), mientras que los más profundos son rojos (>600 km).

El gráfico revela patrones temporales muy marcados:

- **Evento Principal de 2010:** La característica más prominente del gráfico es un evento de gran magnitud (M >8) ocurrido a principios de 2010. Este sismo, de profundidad somera-intermedia (color verde/amarillo), es el de mayor magnitud en todo el catálogo.
- **Secuencia de Réplicas (Aftershocks):** Inmediatamente después del evento principal de 2010, se desata una intensa secuencia de réplicas. Se observa una "nube"vertical de puntos que abarca un rango completo de magnitudes (desde M <2 hasta M >7), indicando una liberación masiva de energía en los meses y años siguientes.
- **Sismicidad de Fondo (Background):** Antes de 2010, la actividad sísmica (sismicidad de fondo) es notablemente menor, con eventos que raramente superan la magnitud 6.
- **Sismicidad Profunda:** Los eventos profundos (puntos rojos, >500 km) son esporádicos y parecen tener magnitudes moderadas (M $\approx$ 4-5). Se observan algunos de estos eventos en los últimos años del catálogo (aprox. 2017-2020).
- **Catálogo Incompleto o Sesgo de Detección:** La densidad de eventos de baja magnitud (M <3) parece aumentar significativamente después de 2010. Esto podría sugerir una mejora en la red de detección sísmica tras el gran terremoto, o simplemente que la secuencia de réplicas continuó produciendo sismos pequeños durante mucho tiempo.

El catálogo está dominado por el terremoto principal de 2010 y su vasta secuencia de réplicas. Este evento singular define la actividad sísmica de las últimas dos décadas en la región, eclipsando la sismicidad de fondo y los eventos profundos esporádicos.

5. Modelamiento

Se entrenaron cinco modelos de clasificación para predecir la variable `type`:

a) Árbol de decisión

b) Random Forest

c) kNN

d) SVM

e) Red neuronal

5.1. Resultados de Evaluación

Modelo	Accuracy
Decision Tree	0.85
Random Forest	0.92
kNN	0.88
SVM	0.87
Neural Net	0.90

Cuadro 3: Accuracy de cada modelo en datos de test.

Se observó que el Random Forest presentó el mejor desempeño, probablemente debido a su capacidad de capturar relaciones no lineales entre variables continuas y categóricas.

6. Reproducción de Resultados

Todos los pasos descritos, que incluyen la descarga, limpieza, visualización y modelamiento, están automatizados en el script `run_analysis.sh`. Para ejecutar el análisis completo, utilice el siguiente comando:

```
bash run_analysis.sh
```

Este script genera automáticamente todos los archivos CSV, figuras y resultados de los modelos, permitiendo reproducir íntegramente el análisis.

7. Conclusiones

- Se realizó un análisis completo de los datos sísmicos, abarcando desde la obtención hasta el modelamiento predictivo.
- La limpieza y transformación de los datos permitió eliminar valores faltantes y estandarizar las columnas de texto.
- Las visualizaciones muestran de manera clara la distribución de magnitudes, profundidades y epicentros geográficos.
- El modelo Random Forest presentó la mayor precisión para predecir el tipo de evento sísmico.
- El informe junto con los scripts permiten reproducir el análisis de manera íntegra y sistemática.