**Título del TFM**

Trabajo Fin de Máster

02 de junio de 2023

# Introducción

## Introducción al proyecto

Los terremotos son vibraciones de la superficie producidas por sacudidas bruscas y repentinas de la corteza terrestre. Comúnmente, son producidos por la actividad de fallas geológicas, otros fenómenos naturales e incluso por el ser humano a través de detonaciones nucleares. Independientemente de su origen, un terremoto puede causar la vibración de la superficie terrestre, así como otros riesgos naturales asociados tales como incendios, tsunamis o incluso corrimientos de tierras que causan la destrucción de vidas humanas y la infraestructura del lugar.

A través del conjunto de datos proporcionado por el Centro Nacional de Información de Terremotos (NEIC, por sus siglas en inglés), organismo que depende del Servicio Geológico de Estados Unidos. La información ha sido proporcionada a través de la plataforma Kaggle, de forma que se pone a disposición del usuario los datos de seísmos producidos en todo el mundo, aportando datos sobre el propio evento sísmico como de su forma de medición. Este reporte concreto se centra en el conjunto eventos sísmicos reportados durante el período de 1965-2016 y con una magnitud igual o superior a 5.5 grados.

Este análisis se centrará en analizar diferentes métodos de machine learning que posibiliten la **predicción de la magnitud de un terremoto** a partir de los datos proporcionados por la plataforma Kaggle.

## Objetivos

Esto el último día

## Estructura

La presente memoria está compuesta por X capítulos, que se detallan a continuación:

**CAPÍTULO 1: Introducción:**

Breve introducción al problema que se aborda a lo largo del trabajo, así como los objetivos propuestos a conseguir y la estructura que se seguirá a lo largo de la memoria descriptiva.

**CAPÍTULO 2: Análisis Descriptivo**

Investigación de los datos que componen el dataset para descubrir patrones, anomalías y evaluar la calidad del propio dato a través de representaciones gráficas.

**CAPÍTULO 3: Transformaciones**

Aplicación de transformaciones al dato para conseguir una mejor modelización posterior.

**CAPÍTULO 4: Experimentación de Modelos de Predicción**

Desarrollo de modelos de predicción aplicando distintas técnicas de Machine Learning y Deep Learning. Tras la experimentación, se muestra un estudio de los resultados obtenidos para cada tipo de modelo, así como una discusión de los resultados obtenidos y su interpretabilidad.

**CAPÍTULO 5: Conclusiones**

Exposición de las conclusiones obtenidas a raíz de las fases anteriores.

**CAPÍTULO 6: Futuros trabajos**

Exposición de posibles aspectos a considerar para continuar el trabajo realizado.

Además, esta memoria se complementa con el documento de Anexos, donde se desarrolla en profundidad la toma de decisiones llevada a cabo durante todo el proceso de análisis de este dataset.

# Análisis Descriptivo

## Composición del dataset

Tal y como se ha introducido brevemente en el apartado anterior, el dataset analizado en este trabajo, aporta datos sobre diferentes características de eventos sísmicos producidos en todo el mundo. Los datos son recopilados a través de una extensa red sismográfica global y gracias a la cooperación internacional de diversas agencias involucradas.

Las variables del dataset original pueden dividirse en **dos subclases diferentes atendiendo al tipo de información que aportan**. En primer lugar, están las variables que aportan información del seísmo en sí, como podrían ser su localización y profundidad a la que se registró el evento. Por otra parte, existe otra subclase de variables que aportan información sobre la forma de medición y errores considerados en el registro, como pueden ser el número de estaciones sísmicas utilizadas para calcular su localización o el algoritmo utilizado para calcular su magnitud. Una de las variables originales fue descartada de inicio, pues estaba compuesta completamente de valores *nulls,* tratándose de un error en los datos de entrada.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Tras una primera investigación del tipo de información que almacenaba la variable, se tomaron algunas **medidas para readecuar su tipo**. En concreto, se utilizaron patrones regex para extraer información relevante de las variables *Time* and *Place*. Además, se pasaron a variables categóricas y a formato fecha aquellas que venían mal en origen. Así, tras una primera depuración y reasignación, la distribución de tipos de variables queda de esta forma:



Una vez reasignado su tipo, se exponen las características principales e insights obtenidos a raíz de su estudio, tanto numérico como visual:

## Distribuciones y representación visual de las variables

En cuanto a las variables numéricas, se ha realizado un análisis numérico y visual a través de histogramas. Se ha observado una **tendencia general a las distribuciones normales con asimetría positiva** de casi todas las variables. En algunos casos se trataba de claros errores en la escritura de datos, pero se ha encontrado una presencia generalizada de valores extremos en las colas derechas de las distribuciones que apuntan a casos extremos.

Imagen que contiene Gráfico

Descripción generada automáticamente Gráfico

Descripción generada automáticamente

Algunas de las variables que no presentan una distribución normal son la variable target Magnitud o la coordenada Longitud.

A pesar de los outliers detectados en las colas derechas de la distribución, **los valores se concentran significativamente en torno a las medianas de las distribuciones**, haciendo que el rango intercuartílico 1Q-3Q sea muy compacto. Las medias de las distribuciones están situadas por lo general a la derecha de la distribución debido a la presencia de outliers.

En cuanto a las variables categóricas, también encontramos un **desbalanceo significativo** en los datos de origen, lo cual indica que se deberán llevar a cabo labores de agrupación en fases posteriores. Algunos de los casos más claros son el de *status* y de *Type*, con más del 99% de valores asignados a la clase principal.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

En otros casos, vemos que el desbalanceo es menor, como ocurre con la red que publica la magnitud del evento sísmico (magSource) o el algoritmo utilizado para calcular la magnitud del suceso (magType).

Gráfico

Descripción generada automáticamente

Por último, uno de los insights que aportan estos gráficos son los el ranking de países más castigados por los eventos sísmicos. La mayor parte de ellos se encuentran en la zona del Cinturón de Fuego del Pacífico, entre la placa tectónica del Pacífico y la Filipina.

Imagen que contiene Interfaz de usuario gráfica

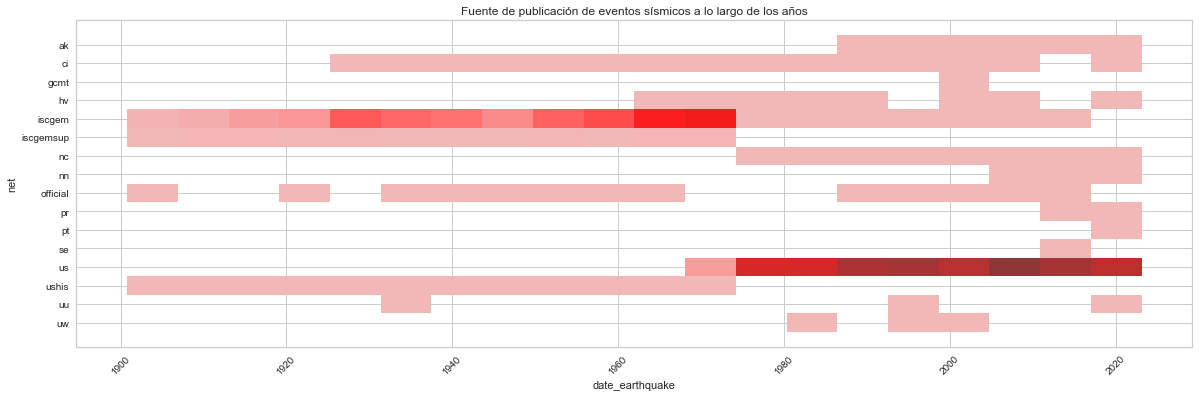
Descripción generada automáticamente

En cuanto a las variables de tipo fecha, vemos **un claro trend de aumento**, llegando sus valores más altos en las recientes décadas. Además, se observa en ambos casos una **tendencia cíclica de aumento** **cada determinado número de años**. Para ampliar la información que aportan estas variables, debería tratarse a través de un estudio de las Series Temporales y eliminar su estacionalidad.

Gráfico, Histograma

Descripción generada automáticamente

Otras gráficas interesantes para comprender la naturaleza de las variables han sido las realizadas con respecto a las fechas de registro de los terremotos. Estas gráficas arrojan insights preliminares que sugieren que, al inicio de esta colección de datos el organismo que más datos aportaba era el ISC-GEM. Sin embargo hoy en día es el organismo americano. Otro insight interesante que aporta la segunda gráfica es que el número de estaciones de medición sísmica utilizadas a lo largo del tiempo ha ido creciendo hasta alcanzar su máximo a mediados de la década de los 2010.



Gráfico

Descripción generada automáticamente

## Valores Missings

El mayor problema que se ha encontrado durante el análisis descriptivo ha sido la cantidad faltante de información. En concreto, las variables que proporcionan información sobre la medición de los seísmos cuentan con una elevada tasa de valores *missings*, lo cual implica que su análisis es prácticamente inviable. Además de sus elevados porcentajes de aparición, desde el 45% al 89% de la información, se han observado patrones de aparición y alta correlación entre los patrones.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Imagen que contiene piano

Descripción generada automáticamente

Finalmente, **se ha optado por descartar estas variables del estudio**. Su alta tasa de valores missing y patrones de aparición sumada a la alta correlación que las variables con gran cantidad de missing tenían entre sí, ha hecho imposible su reconstrucción a partir de otros datos del dataset.

## Representación frente a variable target

Se presentan a continuación diferentes gráficos de las variables con respecto al target Magnitud.

En primer lugar, para una mejor comprensión de la relación entre las coordenadas geográficas y la magnitud, se ha decidido representar esta respecto a Latitud y Longitud sobre un mapamundi. En el gráfico se puede apreciar de forma muy clara la distribución de los seísmos en torno a las uniones de placas tectónicas. Sin embargo, **la magnitud de los seísmos no sigue un patrón espacial aparente**; no se ven zonas localizadas de altas magnitudes, si no que parece un patrón aleatorio.

Mapa de colores

Descripción generada automáticamente con confianza media

Después, se han realizado gráficas 1-1 de las variables predictoras con respecto al objetivo. Se presentan a continuación algunos de los gráficos así como los insights obtenidos.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

**Los organismos gubernamentales que más seísmos detectan son el International Seismological Centre (ISC-GEM) seguido del Instituto Geológico Americano (USGS).** Ambos organismos proporcionan información en un rango de magnitudes muy homogéneo, sin embargo, podemos ver que el USGS proporciona información con **1 decimal** y el ISC-GEM con dos decimales. Ocurre lo mismo con los algoritmos que se utilizan para calcular las magnitudes (**MagType**), que también utiliza distintos niveles de decimales, siendo mw el único que utiliza dos decimales.

Por otra parte, **el tipo de evento sísmico está relacionado con su magnitud**. Las explosiones, tanto nucleares como no nucleares, generan terremotos de baja-media magnitud mientras que los eventos sísmicos provocan magnitudes en todo el rango de la escala Richter.

## Análisis de correlaciones y Relación con variable Target

El análisis de correlaciones de variables muestra una **tendencia general a una baja correlación entre las predictoras numéricas y también con respecto al target**. Las correlaciones más fuertes las encontramos en las variables con una alta tasa de valores *missing* que, como hemos comentado anteriormente, se ha decidido descartar de fases posteriores del estudio.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Este hecho augura una complicada fase de modelización**, puesto que los valores que podemos ver con respecto a la magnitud están en el orden de 0 – 0,1.

A priori, la Latitud, Longitud y Profundidad, no serían capaces de explicar la variabilidad de la Magnitud. Las variable que más correlacionada está es el número de estaciones sísmicas utilizadas en la detección del seísmo (nst), descartada por su alta tasa de valores *missing*.

Por otra parte, **las variables categóricas y de tipo fecha tampoco muestran una relación aparente con respecto a la variable objetivo**. El análisis de V de Crammer arroja resultados negativos de cara a la fase de modelización, indicando que las relaciones son muy bajas con respecto a la variable Magnitud, situándose estas en el rango de 0,02-0,16.

Sin embargo, **se decide que todas las variables pasen a la siguiente fase para poder aplicar transformaciones que puedan maximizar su relación frente a la target.** Además, todas ellas pasan, de forma muy ajustada, el punto de corte marcado por las dos variables aleatorias generadas para el estudio.

Gráfico, Gráfico de barras

Descripción generada automáticamente

**Una de las variables que más relación guarda con la magnitud parece ser la fecha del seísmo, lo cual sería indicativo de la necesidad de analizar la variable a través de Series Temporales.**

# Aplicación de Transformaciones