**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

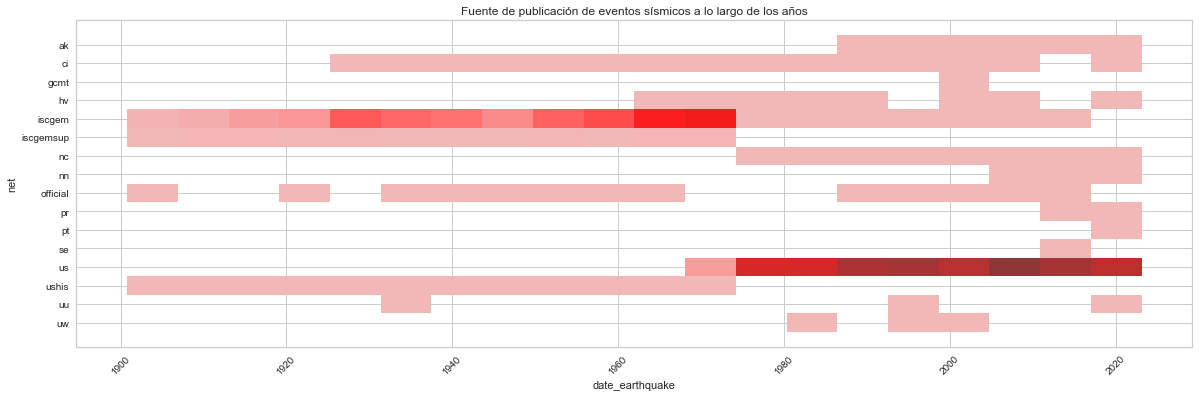
Descripción generada automáticamente

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.**

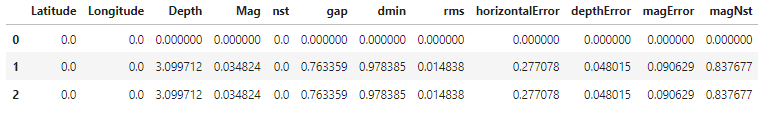
# Transformaciones

## Transformaciones de variables numéricas

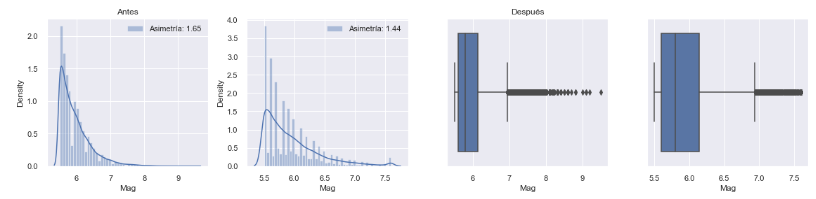
Tal y como hemos podido comprobar en el capítulo anterior, las variables numéricas presentan mayoritariamente una distribución **asimétrica positiva y leptocúrtica**. Sus asimetrías varían en el rango de 0.65 – 77.00, por lo que uno de las primeras transformaciones que se tornan necesarias es la de la gestión de los *outliers* que hemos detectado a través de la representación visual de las variables.

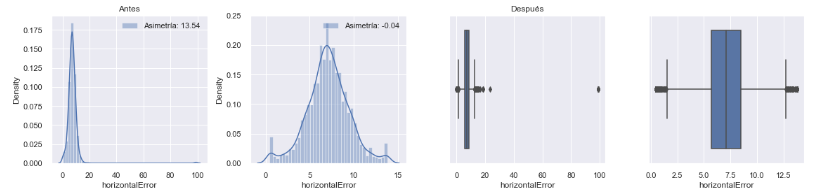
De esta forma, se ha buscado **colapsar las colas derechas de las distribuciones** para poder descartar valores extremos que nos llevaran al desarrollo de modelos poco robustos. Existen dos puntos positivos al analizar la incidencia de outliers:

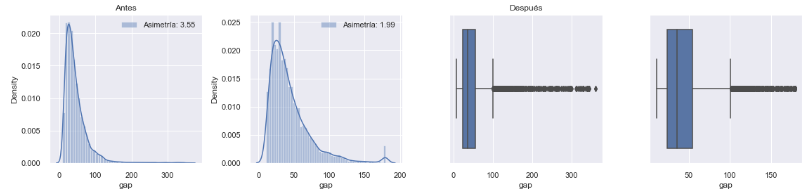
1. No existen outliers en la cola izquierda, únicamente en la derecha.
2. El porcentaje de *outliers* que observamos en las variables apenas llega al 3,1% de los registros, descartando así la presencia de diferentes subpoblaciones.



Se ha optado pues por aplicar una técnica de **winsorización sobre el percentil 1% de la cola derecha.** Al no eliminar los registros y transformarlos en valores extremos aún dentro del rango, se evita distorsionar las medidas de tendencia central y dispersión. Tras aplicar esta técnica de transformación, se dibujan diagramas de boxplot del antes y después. Se observa **de forma general que la asimetría y la curtosis disminuyen** y los diagramas de boxplot se hacen más anchos y se centran en las distribuciones.







Tras mejorar la normalidad de las variables, se ha optado por **aplicar diferentes tipos de transformaciones numéricas** (logarítmicas, potencias y raíces cuadradas y cuartas, exponenciales, …) con el objetivo de conseguir aumentar su linealidad frente a la variable objetivo.

Los análisis numéricos para comparar los resultados de estas transformaciones no arrojan una mejora de la correlación con respecto a la variable target, pues oscila entre 0 y 0,046. **El análisis indica que las variables aportan poco más al modelo que una variable aleatoria.**

Pantalla de computadora

Descripción generada automáticamente

## Transformaciones de variables categóricas

Otra de las transformaciones que se han identificado como necesarias durante el Análisis Descriptivo es la **agrupación de variables categóricas con alta cardinalidad y con categorías subrepresentadas.**

Variables como *Type*, *MagType*, *magSource* o *locationSource* han sido reagrupadas, de forma que las categorías con mayor representación se han dejado intactas y las menos representadas se han agrupado para equilibrar los datos. De esta forma, se reduce el sesgo en el modelo hacia las clases mayoritarias de estas variables.

Se exponen a continuación algunas gráficas de los ejemplos trabajados. En el caso de *MagType\_tr*, se han pasado de 18 a 2 categorías. Algunas de las categorías menos representadas tenían un registro, lo que imposibilitaba el uso de técnicas de árbol en la fase de modelizado. En el caso de *magSource\_tr* nos encontramos el mismo problema, solventándolo en este caso pasando al pasar de 54 a 3 categorías totalmente balanceadas.

Gráfico

Descripción generada automáticamente con confianza baja

Para finalizar el capítulo de transformaciones, se concluye con una gráfica comparativa de la importancia de relaciones a través de V de Crammer, comparando también así la mejora de las variables transformadas frente a las originales.

Podemos ver que, en general, **las transformaciones han mejorado ligeramente su correlación con respecto a la variable Magnitud**. El aspecto negativo es que **el orden de magnitud en el que se mueven las correlaciones es tan pequeño, que están muy cerca de la aleatoriedad**, punto marcado por las variables aleatorias de control.

Finalmente, **se han escogido aquellas transformaciones que presentaban un mayor coeficiente de correlación con respecto a magnitud**. No se puede concluir este capítulo sin la idea de que esta baja correlación general con respecto a la variable objetivo augura unos modelos que no lograrán capturar los patrones subyacentes del dato y cuyo poder predictivo será pues muy limitado.

Gráfico, Gráfico de barras, Gráfico de embudo

Descripción generada automáticamente

# Experimentación de Modelos de Predicción

En esta última fase del estudio para la predicción de la variable Magnitud, se ha trabajado con **Matrices de Diseño** e **Interfaz Fórmula a través del método de mínimos cuadrados ordinarios (OLS),** proporcionadas a través de las librerías de scikit-learn y stats models respectivamente. Se han buscado diferentes matrices de diseño para analizar su ajuste a través del OLS, de forma que se encontraran aquellos coeficientes que mejor ajustaran el modelo a los datos de train y test. El punto de partida ha sido la obtención del **Modelo Completo de Referencia** compuesto por todas las variables predictoras.

Mag ~ status + Type\_tr + MagType\_tr + magSource\_tr + locationSource\_tr + Latitude\_log\_cr + Depth\_sqrt\_cr + Longitude + date\_earthquake + date\_updated

Para este primer Modelo Completo de Referencia se ha utilizado la técnica de regresión lineal básica de LinearRegression. Además, para medir el ajuste del modelo, se han elegido las métricas de **Coeficiente de Determinación** (R2), **Error Absoluto Medio** (MAE) y **Error Cuadrático Medio** (MSE). De esta forma, tenemos métodos de evaluación del ajuste del modelo a los datos observados y errores cometidos en cuanto a la magnitud.

**Los resultados obtenidos han sido muy pobres y están muy lejos de explicar la variación de la variable Magnitud, siendo estos de un R2 de 0,112 y un AIC de 3.249e+04.** Tras eliminar 4 de los predictores que menos importancia relativa tenían con respecto a la variable Magnitud, hemos llegado a un modelo mucho más simple que mantiene esencialmente el mismo R2. Finalmente se conserva la siguiente fórmula:

'Mag ~ + Type\_tr + MagType\_tr + magSource\_tr + locationSource\_tr + date\_earthquake + date\_updated'

Es interesante comentar que, tal y como se ha visto durante todo el informe**, la variable que explicaba más varianza era la fecha en la que se produjo el terremoto (date\_earthquake).**

Se han llevado a cabo diferentes modelos de regresión lineal que podríamos agrupar en las diferentes familias:

Diagrama

Descripción generada automáticamente

En primer lugar, se han creado **modelos de regularización simple**, al ser fácilmente interpretable y nos sirven como punto de partida. Los mejores resultados se han obtenido con el modelo de LinearRegression, pues las regularizaciones Ridge (L1) y Lasso (L2) introducidas en los pesos de los coeficientes no han aportado una mejora en la explicación de la variabilidad de la magnitud del seísmo. Como se puede comprobar, los mejores resultados son aportados por el modelo básico sin regularizaciones de LinearRegression.

Tabla

Descripción generada automáticamente

Después, se ha procedido a probar **algoritmos de árboles de decisión**. Se ha utilizado la técnica de GridSearchCV para ajuste de hiperparámetros, lo que ha permitido encontrar el modelo que más ajustaba a los datos. Se han conseguido R2 del orden de 0.18 – 0.19 y bajar el MAE hasta llegar a 0.295. Por lo general, los árboles que mejor ajustan tienen un número de estimadores alto de 200-300, salvo en el caso del XGBoost, que es de 50. Otro de los elementos comunes es que los bosques tienen una profundidad media-alta de 8 unidades.

Tabla

Descripción generada automáticamente

Como era de esperar, el modelo que más ajusta a los datos es el XGBoost con un R2 de 0.19 y un MAE de 0.29º en la escala Richter.

Finalmente, se ha utilizado el algoritmo **KneighborsRegressor**. Se han ajustado los hiperparámetros para encontrar el número óptimo de vecinos y pesos, pero los resultados no han podido superar a los árboles de decisión.



El R2 obtenido está en torno a 0.14 y su MAE en torno a 0.31. Su rendimiento se sitúa pues entre los modelos de regresión simple y los árboles de decisión.

# Conclusiones

**La primera conclusión a la que llegamos es que la Magnitud de los terremotos no es predecible a partir del conjunto de datos proporcionado.** Tal y como se ha podido ir viendo a lo largo del análisis, las correlaciones de las variables que aportaban información del seísmo apenas tenían correlación con respecto a la magnitud. Esto ha llevado a que los modelos de datos generados no ajustaran a los datos, no pudiendo capturar así una relación significativa entre ambas. Con un resultado final de 19,47% de variabilidad explicada (R2) por el mejor modelo desarrollado, podemos concluir que no es un resultado aceptable y que nos lleva a rechazar la hipótesis nula de que la magnitud es predecible a partir de los datos del conjunto.

Al no encontrar correlaciones significativas con las variables que aportan información sobre el seísmo en sí, **el modelo final se sostiene sobre variables que aportan información de la medición y errores** (Algoritmo de medición de la magnitud, redes que proporcionan la información de la magnitud y la localización del seísmo, …). Por ello, era también predecible el mal resultado de los algoritmos de machine-learning.

Durante el Análisis Exploratorio de Datos, se ha podido observar como las variables de **Latitud** y **Longitud** jugaban un papel fundamental en la ubicación de lugares con gran concentración de terremotos. Tal y como muestra la gráfica, el origen de estos seísmos se origina en los bordes de las placas tectónicas. En concreto, las zonas de mayor actividad sísmica registrada son la del Cinturón de Fuego del Pacífico y la del Cinturón Alpino.

Mapa de colores

Descripción generada automáticamente con confianza media

Sin embargo, **la Latitud y Longitud apenas tienen relación con respecto a la magnitud del evento sísmico**. Tal y como apuntan algunos estudios, la magnitud parece estar más influenciada por factores internos de la Tierra, como la liberación de energía acumulada en las fallas, la rigidez de las rocas o la profundidad del foco sísmico.

Por otra parte, se observa que **la variable que más aportaba a la explicación de la variabilidad de la magnitud es la fecha** en la que se produjo el seísmo. Este hecho implica que es posible que haya una dependencia temporal de la variable con respecto a la target. Sería necesario realizar un análisis más extenso de Series Temporales sobre modelos ARIMA, de Suavizado Exponencial o Redes Neuronales Recurrentes (RNN) para poder eliminar la estacionalidad de los datos y ajustar finalmente modelos.

Con respecto al resto de información que aportaba el dataset, es interesante comentar que **la información de partida no era satisfactoria en cuanto a la cantidad de información faltante**. Desde el primer momento, se tuvieron que desechar 9 columnas de información, casi el 40% del total. La alta tasa de valores *missings* unido a la aparición de patrones y la correlación de las columnas con valores faltantes hizo completamente inviable la reconstrucción del dato.

Otro dato interesante y positivo que añadir en cuanto a la calidad de la información es que los datos que hablaban de errores tenían distribuciones asimétricas positivas, con valores muy concentrados en las zonas bajas de las franjas de errores. Esto nos hace pensar un **dato con una baja tasa de error y actualizado**, puesto que las fechas indicaban que las últimas actualizaciones eran del año 2022.

En cuanto a los modelos que sí han sido analizados en este estudio, **encontramos los mejores resultados en los árboles de decisión**. Este hecho indica que **los datos de partida tenían relaciones no lineales y no monotónicas.** También es interesante apuntar que otra de las razones por las que quizás han funcionado mejor que otros algoritmos es por su atributo de **robustez ante valores atípicos**. Durante la fase de análisis exploratorio de datos, se ha podido comprobar la presencia generalizada de estos valores, sobre todo en la cola derecha de las distribuciones. A pesar de haber realizado una regularización de estos valores extremos a través de técnicas de winsorización, podía verse en las gráficas posteriores que los valores aún presentaban una asimetría positiva que iba a hacer que los modelos de ajuste fueran peores.

**Los** **peores resultados los aportaban la familia de algoritmos de regularización simple**. De hecho, las regularizaciones que aportan los algoritmos de regresión de Ridge y Lasso han empeorado aún más las predicciones realizadas originalmente. Este hecho, sostiene la idea de que los datos de partida no tienen relaciones lineales y no monotónicas con respecto a la magnitud del seísmo.

El mejor resultado se ha obtenido pues a través del algoritmo de **XGBoost**, como cabía esperar. Al introducir técnicas de regularización avanzada en la función de coste, previene el overfitting en los datos. Además, es un algoritmo que funciona bien manejando conjuntos de datos desequilibrados, el cual era adecuado para este análisis.

Finalmente, podemos concluir que los mejores datos proporcionados por el modelo explican el 19.47% de la variabilidad de la magnitud con un error absoluto es de 0.29º en la escala Richter.