# Proyecto final: Entrenamiento y optimización de modelos de Machine Learning

Maximiliano Ramirez

# Abstracto

El proyecto busca generar un modelo de interpretación de datos de terremotos, estos entendidos como movimientos sísmicos de magnitudes significativas y que podrían provocar daños. La motivación radica en la búsqueda de patrones o tendencias que permitan establecer niveles de riesgo geológicos y temporales. Estos niveles de riesgo pueden ayudar a programas de prevención y educación sísmica, e incluso a organismos de emergencia, tanto para actividades preventivas como formativas.

# Resumen de metadatos

La data es obtenida a través de la plataforma kaggle, y se titula Significant Earthquakes 1965-2016, y que incluye movimientos de carácter significativo.

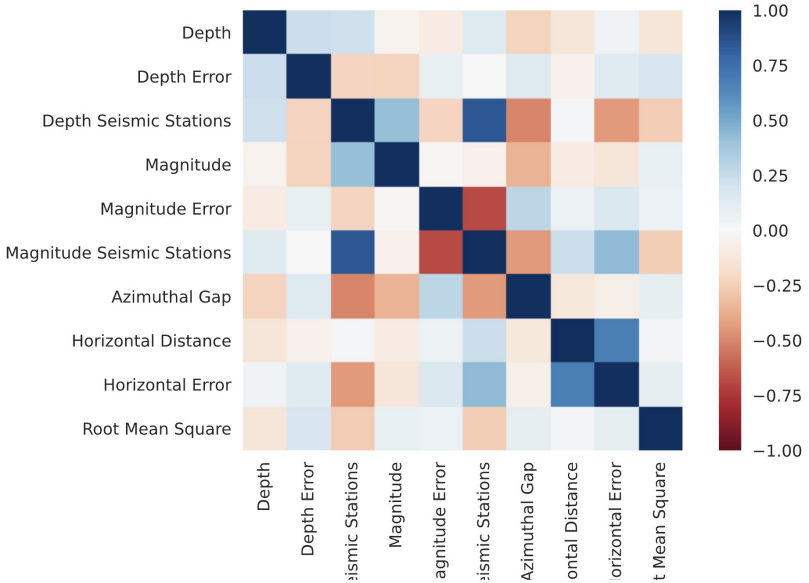
Link: https://www.kaggle.com/datasets/usgs/earthquake-database

```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Date                       23412 non-null  object
 1   Time                       23412 non-null  object
 2   Latitude                   23412 non-null  float64
 3   Longitude                  23412 non-null  float64
 4   Type                       23412 non-null  object
 5   Depth                      23412 non-null  float64
 6   Depth Error                4461 non-null   float64
 7   Depth Seismic Stations     7097 non-null   float64
 8   Magnitude                  23412 non-null  float64
 9   Magnitude Type             23409 non-null  object
 10  Magnitude Error            327 non-null    float64
 11  Magnitude Seismic Stations 2564 non-null   float64
 12  Azimuthal Gap              7299 non-null   float64
 13  Horizontal Distance        1604 non-null   float64
 14  Horizontal Error           1156 non-null   float64
 15  Root Mean Square           17352 non-null  float64
 16  ID                         23412 non-null  object
 17  Source                     23412 non-null  object
 18  Location Source            23412 non-null  object
 19  Magnitude Source           23412 non-null  object
 20  Status                     23412 non-null  object
dtypes: float64(12), object(9)
memory usage: 3.8+ MB
```

# Preguntas e hipótesis

El objetivo es identificar las variables físicas que juegan un rol preponderante en la magnitud de un terremoto. Sean estas la profundidad, latitud y longitud geográficas, entre otras que se registran, como se puede observar en el mapa de calor de correlaciones entre las variables del dataset.

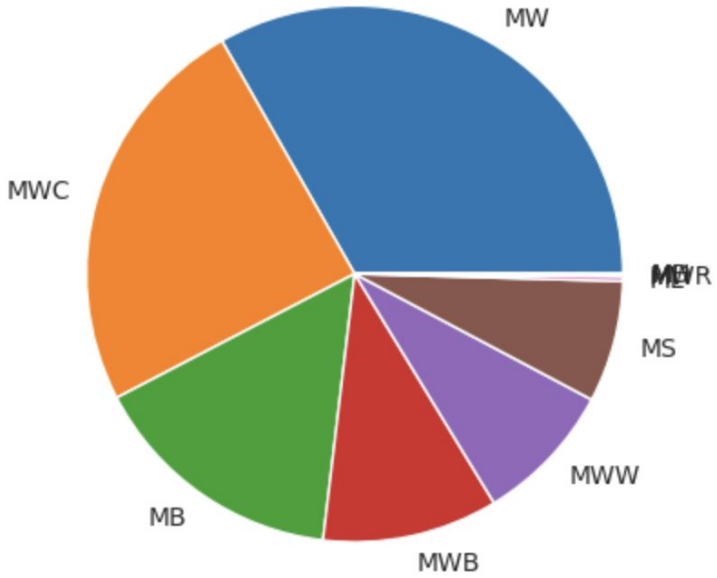
En primer lugar se emplean modelos de machine learning para mejorar la calidad de la base de datos, para luego emplear un modelo de clasificación sobre la data.



# Evaluación de modelos de ML

En algunos casos la información faltante puede ser interpolada o interpretada, en este caso evaluaremos modelos de ML para interpretar datos faltantes.

Primero se identifica como variable objetivo el tipo de magnitud registrada (ver gráfico de distribución de dicha variable), para luego comparar un grupo de tres modelos (RandomForest contra Árbol de decisión y regresión logística).



# Evaluación de modelos de ML

Podemos concluir que Random forest es el modelo con mejores resultados para la data en análisis.

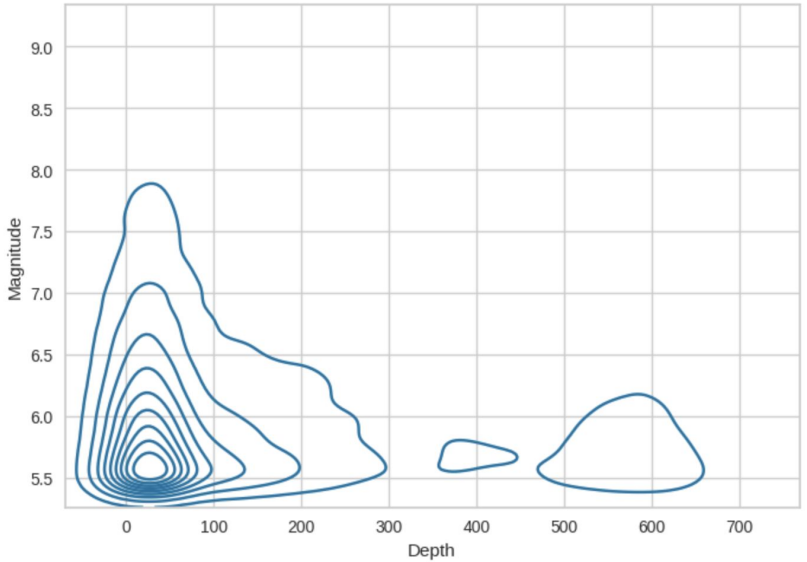
| Modelo | Accuracy | F1-Score |
|---|---|---|
| RandomForest | 0,63 | 0,63 |
| Árbol de decisión | 0,57 | 0,52 |
| Regresión logística | 0,57 | 0,52 |

# La ubicación geográfica

Al observar la data graficada sobre un mapa, obtendremos una imagen familiar, donde las zonas de alta actividad sísmica se hacen notar fuertemente (véase el cinturón de fuego del pacifico). La ubicación de los terremotos está fuertemente asociada a las zonas de confluencia tectónica, caracterizadas por su alta actividad. Fuera de estas zonas la actividad decae fuertemente.

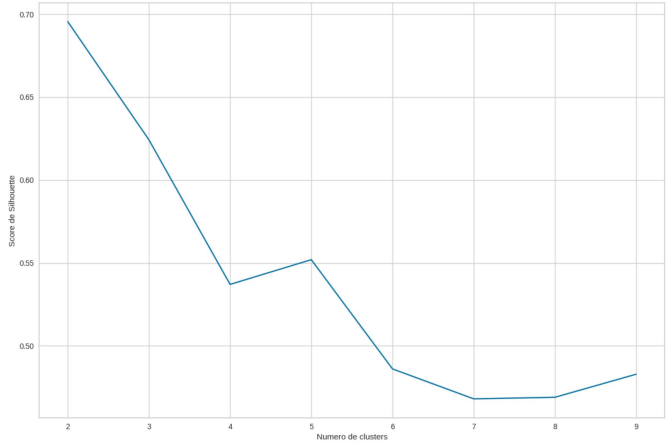
# ¿Relación entre la profundidad y la magnitud?

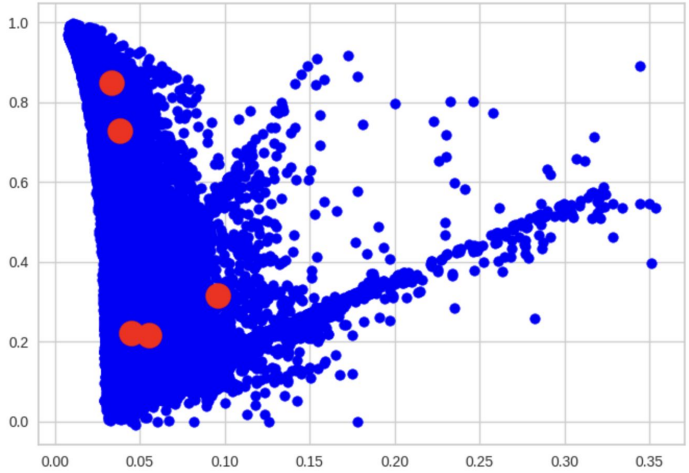
En esta pregunta nos detendremos a analizar los datos con mayor detención, ya que a simple vista no se ve una relación evidente, pero a la vez también se observan zonas con alta densidad eventos, lo que sugiere la existencia de clusters ocultos dentro de los datos.



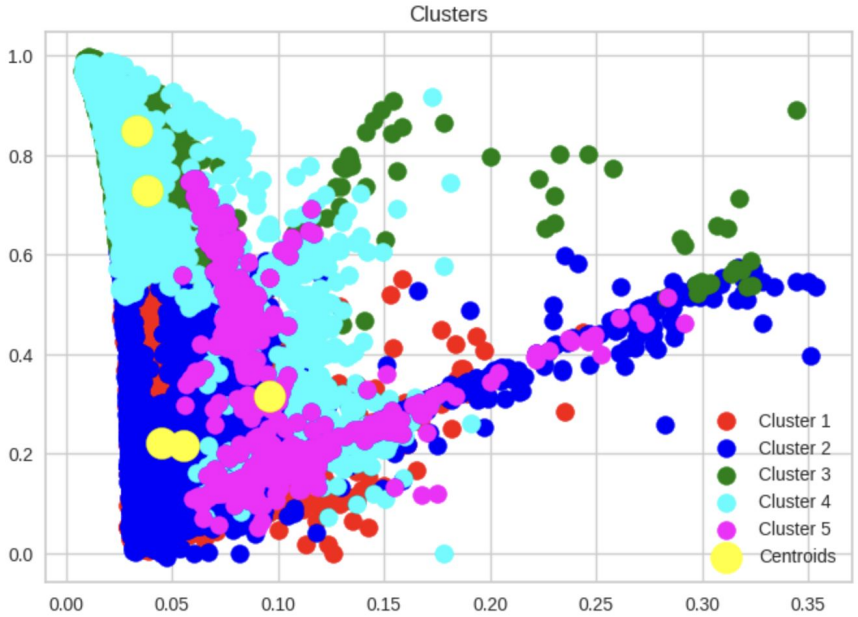
# ¿Relación entre la profundidad y la magnitud?

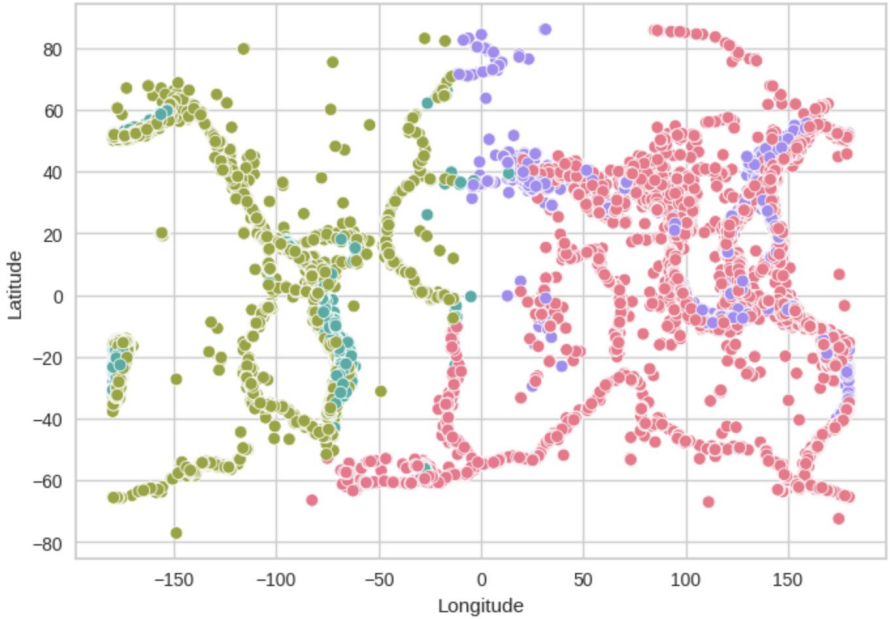
Mediante un modelo de clasificación es posible identificar una serie de clusters de interés para el estudio, que para el caso de los terremotos nos interesa separarlos en regiones geográficas de interés (mayor a 2).



# Información obtenida de los clusters



# Optimización, validación y conclusiones

Con la nueva información obtenida de los clusters, se genera un nuevo modelo de ML sobre el cual se realizarán diversos test tanto de validación (K-folds), hipertunning (RandomizedSearch), y obtención de indicadores para modelo Random Forest.

| Indicador | Random Forest (gini, profundidad=7) |
|---|---|
| MAE | 0,008 |
| MSE | 0,023 |
| RMS | 0,152 |