**Tarea 5**

Judith Carolina Lucio Garza

Natural Disasters

La base de datos con la que se está trabajando se extrae de EM-Dat, con registros históricos de 100 años para distintos desastres naturales alrededor del mundo. El formato se encuentra en csv, ya que, a la fecha, para fines de estas tareas no se ha hallado una API que permita extracción de información histórica respecto al tema.

La idea inicial involucra filtrar para considerar únicamente los terremotos en la base, sin embargo, por temas de errores en formato sin identificar, se utilizan los primeros 100 registros de la base completa una vez que se hizo su pre procesamiento.

Las variables que conforman el dataframe en pySpark son las 22 siguientes:

* magnitude: double (nullable = true)
* start\_year: integer (nullable = true)
* start\_month: integer (nullable = true)
* start\_day: integer (nullable = true)
* end\_year: integer (nullable = true)
* end\_month: integer (nullable = true)
* end\_day: integer (nullable = true)
* total\_deaths: integer (nullable = false)
* total\_affected: integer (nullable = false)
* insured\_damage\_adjusted: integer (nullable = false)
* start\_date: date (nullable = true)
* end\_date: date (nullable = true)
* historic\_indexado: integer (nullable = true)
* disaster\_group\_indexado: integer (nullable = true)
* disaster\_subgroup\_indexado: integer (nullable = true)
* disaster\_type\_indexado: integer (nullable = true)
* disaster\_subtype\_indexado: integer (nullable = true)
* iso\_indexado: integer (nullable = true)
* country\_indexado: integer (nullable = true)
* subregion\_indexado: integer (nullable = true)
* region\_indexado: integer (nullable = true)
* location\_indexado: integer (nullable = true)

Para identificar la cantidad de personas afectadas, se han seleccionado como relevantes las variables magnitude, disaster\_group\_indexado, disaster\_type\_indexado, disaster\_subtype\_indexado y región\_indexado. El glosario para identificar sus índices es el siguiente:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Índice** | **Definición** |
| disaster\_group\_indexado | 1 | Natural |
| disaster\_type\_indexado | 1 | Flood |
| disaster\_type\_indexado | 2 | Earthquake |
| disaster\_type\_indexado | 3 | Storm |
| disaster\_type\_indexado | 4 | Extreme temperature |
| disaster\_type\_indexado | 5 | Wildfire |
| disaster\_type\_indexado | 6 | Epidemic |
| disaster\_subtype\_indexado | 1 | Ground movement |
| disaster\_subtype\_indexado | 2 | Riverine flood |
| disaster\_subtype\_indexado | 3 | Tropical cyclone |
| disaster\_subtype\_indexado | 4 | Flash flood |
| disaster\_subtype\_indexado | 5 | Flood (General) |
| disaster\_subtype\_indexado | 6 | Storm (General) |
| disaster\_subtype\_indexado | 7 | Forest fire |
| disaster\_subtype\_indexado | 8 | Cold wave |
| disaster\_subtype\_indexado | 9 | Heat wave |
| disaster\_subtype\_indexado | 10 | Tornado |
| disaster\_subtype\_indexado | 11 | Extra-tropical storm |
| disaster\_subtype\_indexado | 12 | Tsunami |
| disaster\_subtype\_indexado | 13 | Severe weather |
| disaster\_subtype\_indexado | 14 | Coastal flood |
| disaster\_subtype\_indexado | 15 | Land fire (Brush, Bush, Pasture) |
| disaster\_subtype\_indexado | 16 | Lightning/Thunderstorms |
| disaster\_subtype\_indexado | 17 | Blizzard/Winter storm |
| disaster\_subtype\_indexado | 18 | Severe winter conditions |
| disaster\_subtype\_indexado | 19 | Viral disease |
| disaster\_subtype\_indexado | 20 | Hail |
| disaster\_subtype\_indexado | 21 | Wildfire (General) |
| disaster\_subtype\_indexado | 22 | Sand/Dust storm |
| disaster\_subtype\_indexado | 23 | Storm surge |
| region\_indexado | 1 | Asia |
| region\_indexado | 2 | Americas |
| region\_indexado | 3 | Europe |
| region\_indexado | 4 | Africa |
| region\_indexado | 5 | Oceania |

A través de la selección y estandarización de las principales características, se tomó el 70% de los datos como datos de entrenamiento para el modelo de regresión logística. Se hizo la prueba en el 30% restante, la cual dio un nivel de precisión del 56.1%, el cual es deficiente.

Continuando con un modelo de regresión lineal, para los mismos 100 datos previamente seleccionados, así como las características consideradas relevantes, se tuvo un desempeño por R ajustada del -0.05%. El error cuadrado medio es de 27096.57, mostrando gran variabilidad. Esto indica que el modelo tiene un desempeño muy deficiente para predecir el número de personas afectadas por los desastres.