

## **Presentacion Proyecto**

**Introducción a la gestión de datos Geocientíficos**

**Juan Pablo Lozada Escobar**

**Docente Ever Herrera Rios**

**Universidad Nacional de Colombia – Sede Medellin**

**Facultad de Minas**

**Diciembre de 2024**

## Contenido

<u>Introducción</u>	<u>3</u>
<u>Generalidades del Dataset</u>	<u>3</u>
<u>Limpieza de los datos</u>	<u>7</u>
<u>Estadísticos del conjunto de datos y de las variables involucradas</u>	<u>7</u>
<u>Presentación general de los modelos</u>	<u>8</u>
<u>Comparación del rendimiento entre modelos</u>	<u>8</u>

# **Introducción**

En el estudio de los procesos de inestabilidad del terreno, los inventarios de movimientos en masa se han consolidado como una herramienta fundamental para comprender la ocurrencia, distribución y caracterización de fenómenos como deslizamientos, flujos y caídas de roca.

La ciencia ha mostrado que la calidad y detalle de los datos disponibles son determinantes para el desarrollo de modelos confiables de análisis espacial, evaluación de amenazas y estimación de riesgo (Guzzetti et al., 2012; SGC, 2020). A lo largo de las últimas décadas, el uso de datasets geoespaciales estandarizados ha permitido avanzar en metodologías de clasificación automática, análisis multivariado, modelación de susceptibilidad y, más recientemente, el entrenamiento de algoritmos basados en aprendizaje automático para identificar patrones asociados a los movimientos en masa.

Sin embargo, antes de aplicar cualquier técnica de modelamiento, es indispensable realizar una revisión sistemática del dataset que incluye su origen, estructura, calidad y consistencia interna. Los problemas típicos documentados—como codificaciones heterogéneas, valores faltantes, errores topológicos o atributos incompletos—pueden afectar significativamente la interpretación y el rendimiento de los modelos. Por ello, este trabajo inicia con una revisión detallada de las características del inventario del Servicio Geológico Colombiano, explorando sus variables, su estructura temática y su utilidad dentro de un Sistema de Información Geográfica (SIG).

A partir de esta comprensión del contexto del dataset, el análisis se orienta hacia la pregunta central de negocio: ¿cómo puede el tratamiento, limpieza y modelamiento estadístico de este inventario mejorar la capacidad de clasificación o predicción de los tipos y subtipos de movimientos en masa, y qué modelo ofrece el mejor desempeño según las métricas evaluadas?

Este trabajo tiene por propósito evaluar, mediante técnicas estadísticas y modelos computacionales, la calidad de la información disponible y su potencial para apoyar la toma de decisiones en escenarios de gestión del riesgo.

## **Generalidades del Dataset**

Explicación técnica de cada columna del dataset geoespacial

El dataset corresponde a un inventario de movimientos en masa y procesos gravitacionales, estructurado para ser utilizado dentro de un Sistema de Información Geográfica (SIG). (SGC, 2020).

Cada columna representa un atributo que describe características geomorfológicas, administrativas o cartográficas de los eventos registrados.

## **1. IDENTIFICACIÓN DEL OBJETO**

Es un identificador geométrico único asignado automáticamente por el sistema o la geodatabase.

Generalmente corresponde al GlobalID o un campo de identificación extendido que combina la ubicación dentro del dataset.

Su función principal es permitir:

Trazabilidad del objeto espacial.

Integridad en procesos de edición.

Uniones internas con otras tablas relacionadas del SIG.

## **2. IDENTIFICACIÓN**

Representa un ID numérico secuencial asignado a cada movimiento registrado.

A diferencia del campo anterior, este ID es más manejable y puede ser utilizado para:

- Referencia rápida en análisis y reportes
- Relación con información de campo.
- Organización de registros dentro del inventario.

## **3. Inventario\_Movimiento**

Corresponde a un valor cuantitativo perteneciente al inventario base, posiblemente asociado a:

- Un código interno de clasificación,
- Un índice de registro,
- O una medida derivada del inventario general (frecuencia, ordenamiento, etc.).

Su significado preciso depende del modelo institucional del inventario, pero suele utilizarse para procesos de validación y depuración de datos.

## **4. F35DOV\_TIP**

Es un código de dominio o tipología institucional estandarizada.

Permite homogenizar los movimientos según catálogos corporativos o normas técnicas.

- Facilita la interoperabilidad con:
- Otros sistemas corporativos,
- Servicios web,
- Modelos de datos oficiales (por ejemplo, para entidades ambientales o de gestión del riesgo).

Sus valores corresponden a:

67 = Caída

68 = Volcamiento

69 = Deslizamiento

70 = Propagación lateral

71 = Flujo

72 = Reptación

73 = Deformaciones gravitacionales

Esta etiqueta está directamente relacionada con Tipo\_Movimiento (se puede considerar a F35DOV\_TIP como el identificador del tipo de movimiento registrado).

## 5. Tipo\_Movimiento

Describe la categoría principal del proceso de movimiento en masa, como:

- Deslizamiento
- Flujo
- Caída
- Deformaciones gravitacionales
- Propagación lateral
- Reptación
- Volcamiento

Este campo representa la primera jerarquía clasificatoria del fenómeno, usualmente asociada a criterios mecánicos y geomorfológicos.

Su importancia radica en que:

Es la variable base para los análisis comparativos entre tipologías. Permite agrupar eventos para estudios estadísticos.

Es el eje de la simbología temática en el mapa.

## 6. Subtipo\_Movimiento

Es un código numérico que representa el subtipo específico del movimiento.

Utiliza un sistema de codificación interno para distinguir variaciones dentro de un mismo tipo de proceso.

Ejemplos según la tabla:

1 = Volcamiento flexural de roca

2 = Avalanche de rocas

4 = Deslizamiento translacional

3 = Deslizamiento rotacional

5 = Flujo de detritos

6 = Flujo de lodo

7 = Propagacion lateral lenta

8 = Reptacion de suelos

9 = Solifluxion

10 = Volcamiento de roca

12 = Deslizamiento por flujo

14 = Deformaciones gravitacionales profundas

15 = Caída de roca

16 = Avalanche de detritos

17 = Flujo de tierra

18 = Deslizamiento de cuña

19 = Propagacion lateral por licuación

20 = Caida de detritos

23 = Caida de tierras

24 = Volcamiento del macizo rocoso

25 = Crecida de detritos

81 = Deslizamiento translacional en cuña

82 = Deslizamiento translacional planar

83 = Deslizamiento por licuación

Este enfoque codificado permite:

- Reducir errores en bases de datos,
- Agilizar filtros,
- Mantener estandarización en procesos de análisis automatizado.

## **7. Subtipo\_nombre**

Es la descripción literal del subtipo de movimiento.

A diferencia del código numérico anterior, este campo es plenamente legible para el usuario final.

Ejemplos:

- “Deslizamiento translacional”
- “Deslizamiento rotacional”
- “Flujo de lodo”
- “Caída de roca”
- Uso principal:
- Claridad interpretativa,
- Presentación en informes,
- Etiquetado en layouts cartográficos.

## **8. Etiqueta**

Es un rótulo simplificado que resume el tipo o subtipo del movimiento. Se utiliza principalmente para:

- Etiquetado automático en el mapa,
- Identificación rápida de entidades
- Representaciones simplificadas en visores geográficos web.

Ejemplos del dataset:

“Deslizamiento”

“Deslizamiento rotacional”

“Flujo”

“Caída”

## **9. Representación\_mapa**

Campo especializado destinado a definir la simbología o regla de representación cartográfica.

Contiene códigos internos o abreviaturas utilizadas por el SIG para aplicar simbología.

Ejemplos observados:

dt → Deslizamiento translacional

dr. → Deslizamiento rotacional

Florida → Sobresimbología para flujos (probablemente color o textura)

cr → Caída de roca

Su relevancia reside en:

Garantizar consistencia visual,

Permitir simbologías automatizadas,

Mantener estandarización entre productos cartográficos oficiales.

## **10. ESRI\_OID**

Identificador único generado automáticamente por software ESRI (ArcGIS). Es utilizado para:

- Control interno del dataset,
- Uniones temporales,
- Administración de objetos GIS.

No debe ser modificado manualmente, pues es parte integral de la estructura de la geodatabase.

## **Justificación del uso del dataset en el Sistema de Información Geográfica**

El presente dataset constituye un insumo fundamental para el análisis, gestión y representación espacial de los procesos de inestabilidad del terreno. Su estructura de atributos ha sido diseñada para permitir una clasificación estandarizada y jerárquica de los diferentes tipos y subtipos de movimiento, lo cual resulta indispensable para comprender la

dinámica geomorfológica de un área y para desarrollar modelos confiables de evaluación de amenazas y riesgos.

La incorporación de campos como Tipo\_Movimiento, Subtipo\_Movimiento e Inventory\_Movimiento posibilita una caracterización precisa del fenómeno, en consonancia con sistemas de clasificación ampliamente aceptados en la literatura técnica (por ejemplo, las tipologías de Varnes y modificaciones posteriores). Esta estructura temática permite no solo diferenciar los procesos según su naturaleza mecánica, sino también identificar patrones espaciales y temporales relevantes para la gestión del territorio.

Asimismo, la presencia de atributos orientados a la representación cartográfica, como Representación\_mapa y Etiqueta, facilita la estandarización visual de la información en productos cartográficos oficiales y contribuye a garantizar la legibilidad, consistencia y comparabilidad entre capas temáticas. Estos componentes tienen un valor significativo en la comunicación de resultados, ya que mejoran la interpretación por parte de usuarios técnicos y no técnicos.

Por otro lado, los identificadores únicos (ESRI\_OID e ID) permiten establecer relaciones robustas con otras bases de datos, metadatos y registros administrativos, favoreciendo la interoperabilidad y el mantenimiento de la integridad del sistema de información. La estructura del dataset responde a los principios de organización de datos espaciales definidos por normas internacionales como ISO 19115 e ISO 19110, lo cual contribuye a garantizar la trazabilidad, la calidad y la consistencia del inventario.

En suma, el uso de este dataset dentro del SIG se justifica por su capacidad para integrar información temática, geomorfológica y administrativa en un único modelo coherente, facilitando análisis multiescalares, evaluaciones comparativas y procesos de toma de decisiones informados. La estandarización de sus atributos permite no solo representar adecuadamente los fenómenos asociados a movimientos en masa, sino también asegurar que los resultados derivados del análisis espacial sean reproducibles, verificables y aplicables en contextos de planificación territorial, gestión del riesgo y estudios científicos.

## Limpieza de los datos

Al analizar cada una de las columnas de la base de datos, vemos que es necesario eliminar la columna F35DOV\_TIP y Subtipo\_Movimiento pues ambas son redundante con las columnas Tipo\_Movimiento y Subtipo\_nombre, respectivamente (F35DOV\_TIP da un numero para clasificar cada Tipo\_Movimiento, y Subtipo\_Movimiento da un numero para clasificar cada Subtipo\_nombre).

También eliminaremos la columna Representacion\_mapa pues esta corresponde a una etiqueta mnemotécnica que representa la misma información que existe en Subtipo\_nombre (Representacion\_mapa corresponde al tipo de polígono que se uso para representar en el mapa de ArcGis).

Un punto importante durante la limpieza de la base de datos fue estandarización de categorías porque, en datasets institucionales, los nombres pueden venir con:

- Mayúsculas/minúsculas mezcladas
- Tildes inconsistentes
- Espacios extras
- Variaciones mínimas que afectan el modelamiento

Con esta limpieza obtenemos:

	Columna	Tipo_de_Dato	Valores_Nulos	Valores_Únicos
	<b>FID</b>	FID	int64	0 6826
	<b>OBJECTID</b>	OBJECTID	float64	0 4
	<b>ID</b>	ID	int64	0 6826
<b>Inventario_Movimiento</b>	Inventario_Movimiento		int64	0 6826
<b>Tipo_Movimiento</b>	Tipo_Movimiento		object	0 7
<b>Subtipo_nombre</b>	Subtipo_nombre		object	0 24
<b>Etiqueta</b>	Etiqueta		object	0 10
<b>ESRI_OID</b>	ESRI_OID		int64	0 6826
<b>x</b>	x		float64	0 6060
<b>y</b>	y		float64	0 6177

# Estadísticos del conjunto de datos y de las variables involucradas

En vista que las variables numéricas son identificadores pasamos a revisar las estadísticas de las variables categóricas

```
===== TIPO_MOVIMIENTO =====
Tipo_Movimiento
deslizamiento          3900
caida                  1196
flujo                  1064
reptacion               595
volcamiento              59
deformaciones gravitacionales   8
propagacion lateral           4
Name: count, dtype: int64

===== SUBTIPO_NOMBRE =====
Subtipo_nombre
deslizamiento traslacional      2130
deslizamiento rotacional       1475
caida de detritos             547
reptacion de suelos            515
caida de roca                 464
flujo de detritos              423
flujo de lodo                  257
flujo de tierra                219
deslizamiento traslacional planar 218
caida de tierras               185
Name: count, dtype: int64

===== ETIQUETA =====
Etiqueta
deslizamiento          2420
deslizamiento rotacional 1475
caida                  1196
flujo                  951
reptacion               595
avalancha               113
volcamiento              59
deformacion gravitacional profunda 8
deslizamiento por licuacion    5
propagacion lateral           4
Name: count, dtype: int64
```

De cada etiqueta en cada categoría podemos analizar sus respectivas frecuencias relativas y absolutas:

Tipo_Movimiento	Frecuencia Absoluta	Frecuencia Relativa
deslizamiento	3900	0.571
caida	1196	0.175
flujo	1064	0.156
reptacion	595	0.087
volcamiento	59	0.009
deformaciones gravitacionales	8	0.001
propagacion lateral	4	0.001

===== SUBTIPO\_NOMBRE =====

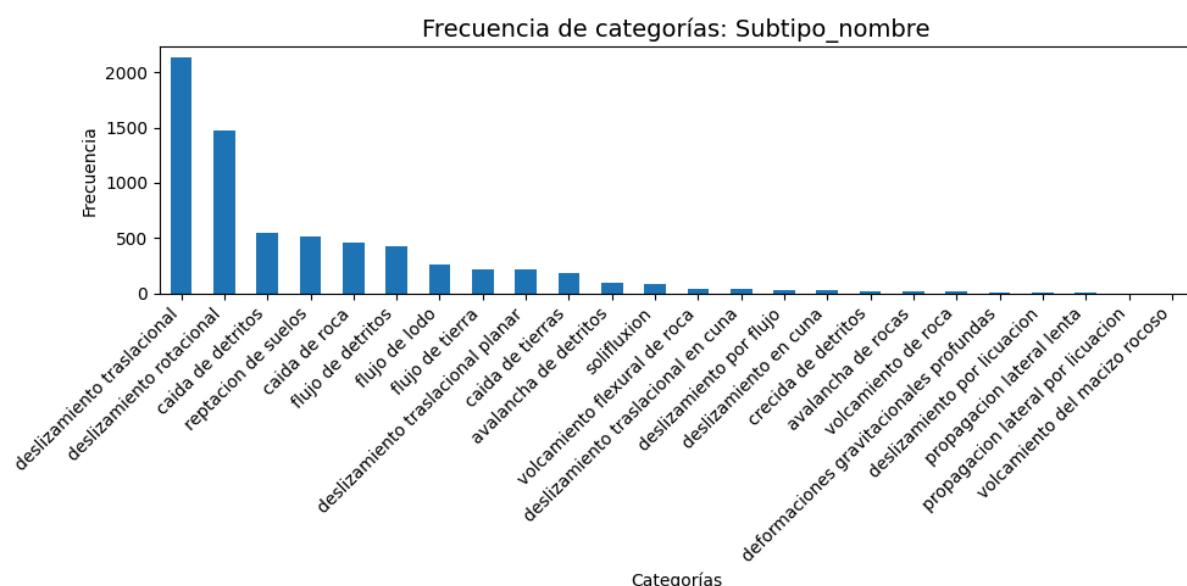
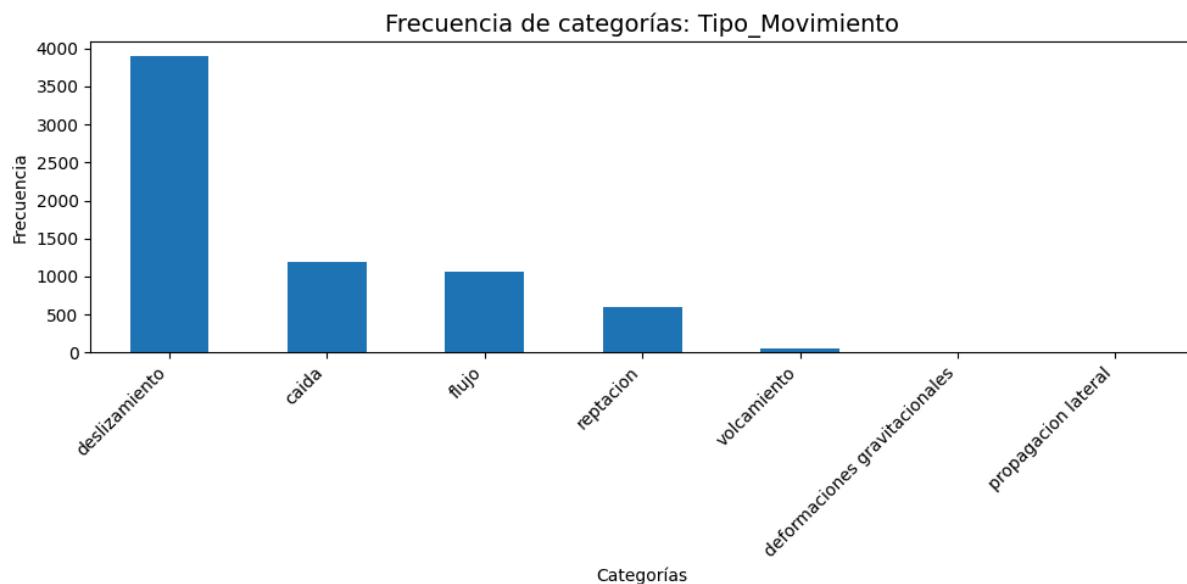
	Frecuencia Absoluta	Frecuencia Relativa
<b>Subtipo_nombre</b>		
<b>deslizamiento traslacional</b>	2130	0.312
<b>deslizamiento rotacional</b>	1475	0.216
<b>caida de detritos</b>	547	0.080
<b>reptacion de suelos</b>	515	0.075
<b>caida de roca</b>	464	0.068
<b>flujo de detritos</b>	423	0.062
<b>flujo de lodo</b>	257	0.038
<b>flujo de tierra</b>	219	0.032
<b>deslizamiento traslacional planar</b>	218	0.032
<b>caida de tierras</b>	185	0.027
<b>avalancha de detritos</b>	97	0.014
<b>solifluxion</b>	80	0.012
<b>volcamiento flexural de roca</b>	45	0.007
<b>deslizamiento traslacional en cuna</b>	43	0.006
<b>deslizamiento por flujo</b>	32	0.005
<b>deslizamiento en cuna</b>	29	0.004
<b>crecida de detritos</b>	20	0.003
<b>avalancha de rocas</b>	16	0.002
<b>volcamiento de roca</b>	13	0.002
<b>deformaciones gravitacionales profundas</b>	8	0.001
<b>deslizamiento por licuacion</b>	5	0.001
<b>propagacion lateral lenta</b>	3	0.000
<b>propagacion lateral por licuacion</b>	1	0.000
<b>volcamiento del macizo rocoso</b>	1	0.000

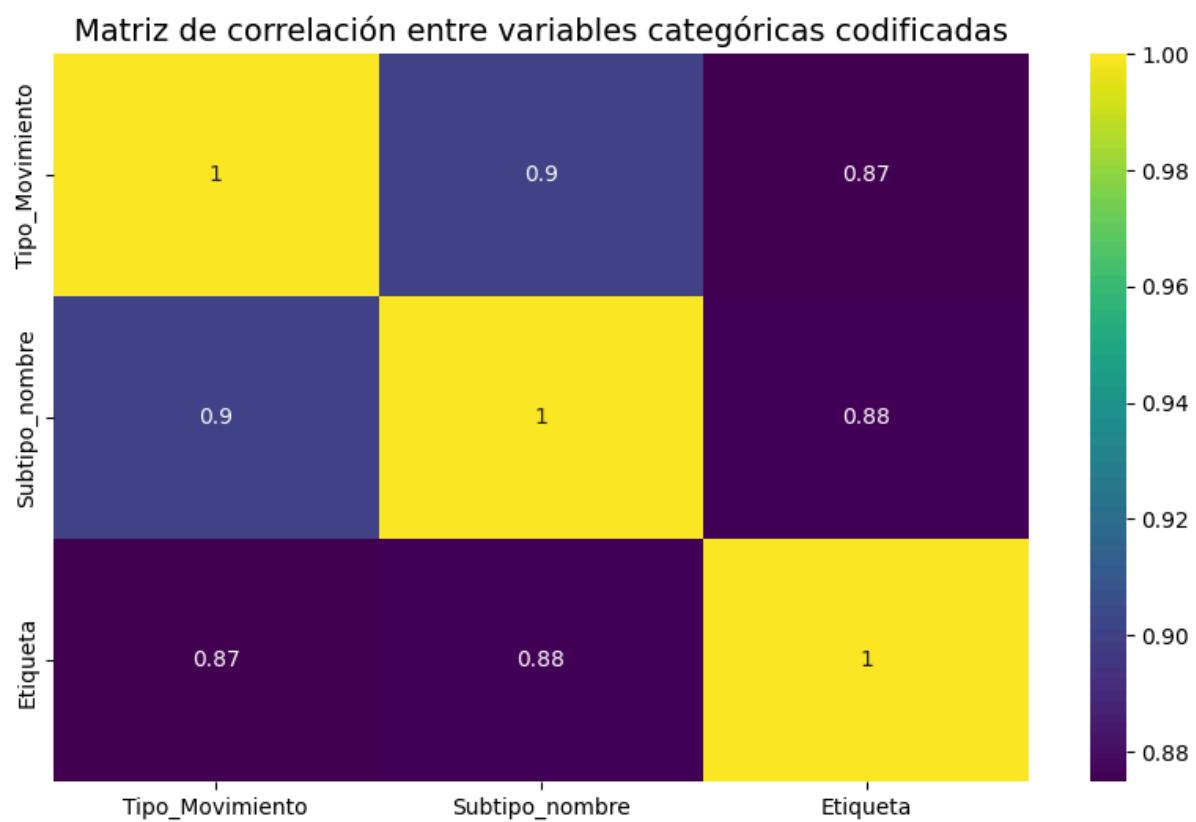
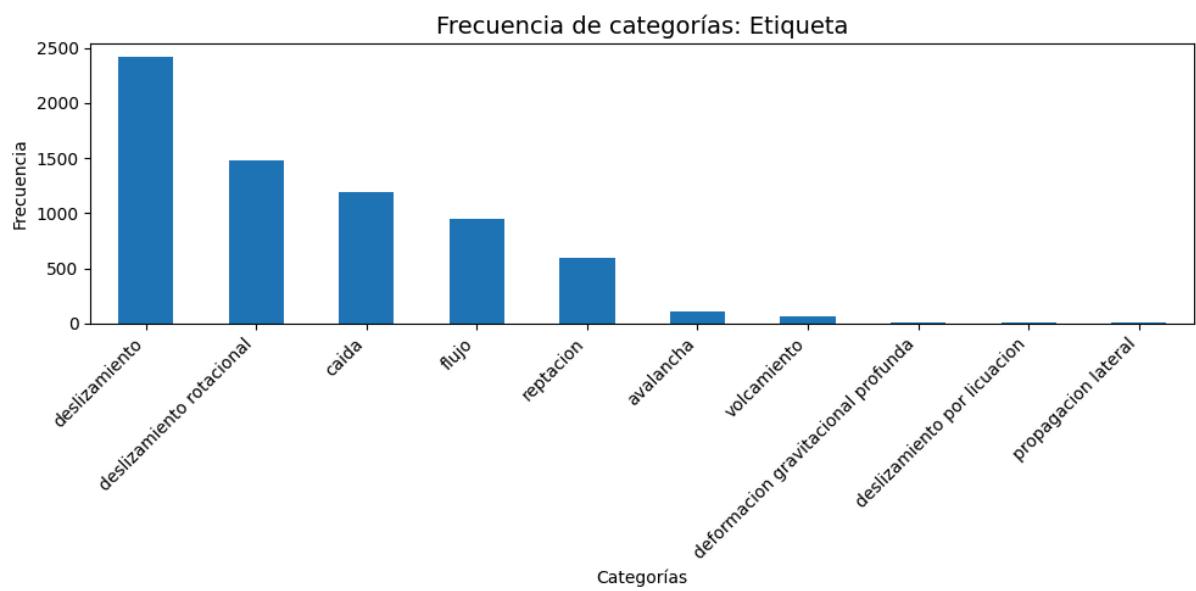
===== ETIQUETA =====

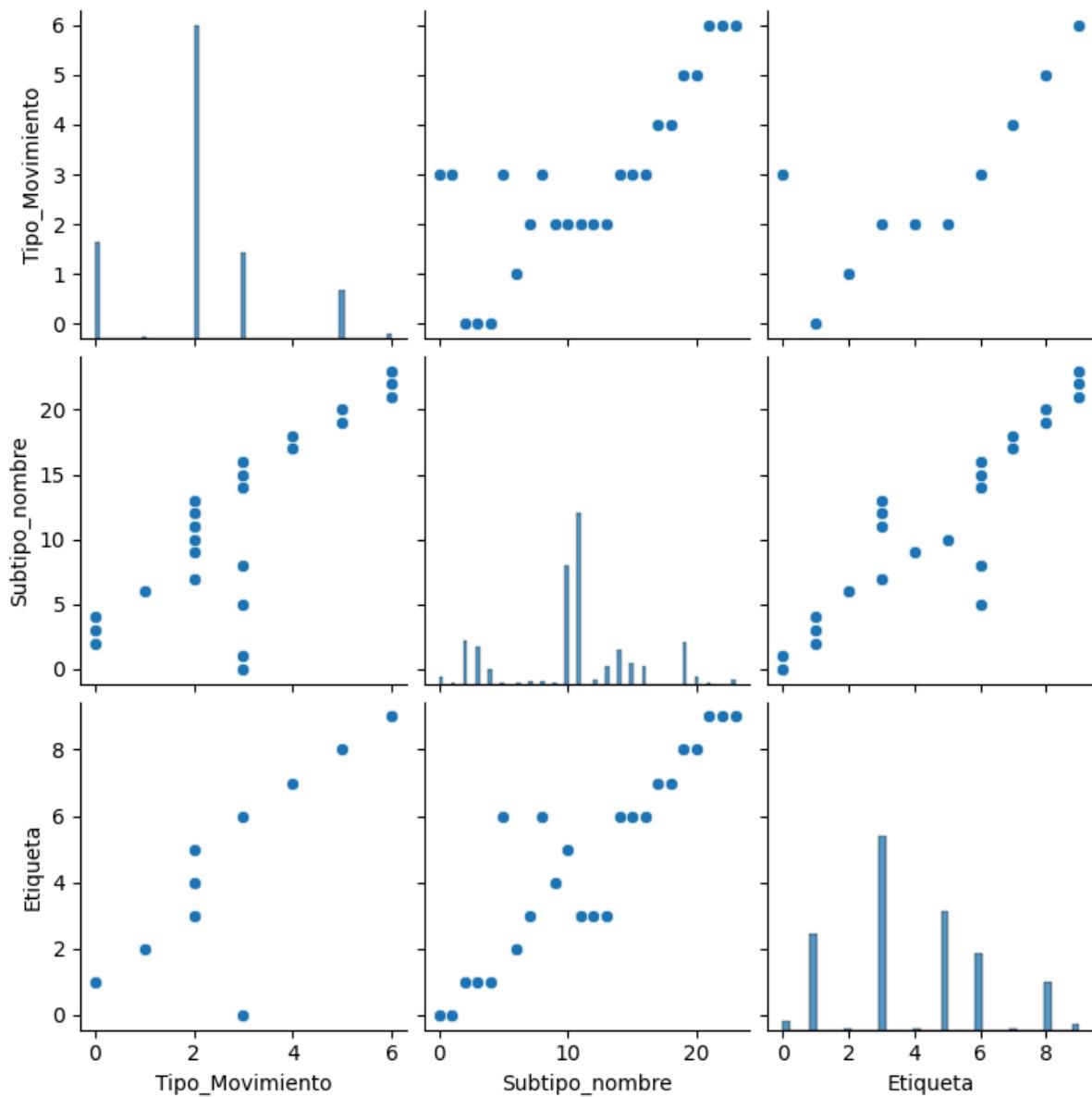
	Frecuencia Absoluta	Frecuencia Relativa
<b>Etiqueta</b>		
<b>deslizamiento</b>	2420	0.355
<b>deslizamiento rotacional</b>	1475	0.216
<b>caida</b>	1196	0.175
<b>flujo</b>	951	0.139
<b>reptacion</b>	595	0.087
<b>avalancha</b>	113	0.017
<b>volcamiento</b>	59	0.009
<b>deformacion gravitacional profunda</b>	8	0.001
<b>deslizamiento por licuacion</b>	5	0.001
<b>propagacion lateral</b>	4	0.001

Para las variables categóricas del inventario, se calcularon frecuencias absolutas y relativas con el fin de identificar la distribución de los tipos y subtipos de movimiento, así como de las etiquetas y categorías administrativas asociadas. Los resultados muestran que las categorías “deslizamiento” y “deslizamiento translacional” son las más representativas dentro del conjunto de datos.

Adicionalmente, se realizaron visualizaciones mediante gráficos de barras para cada variable categórica, lo que permitió observar la concentración de registros en determinadas clases. Finalmente, se generó un heatmap basado en codificación numérica de las categorías, con el objetivo de explorar posibles patrones de co-ocurrencia entre variables temáticas del inventario.







Al analizar la correlación entre variables vemos que existe una fuerte correlación entre:

- Etiqueta-Subtipo\_nombre
- Etiqueta-Tipo\_movimiento
- Subtipo\_nombre-Tipo\_movimiento

# Presentación general de los modelos

Al ejecutar el modelo soportado en el cuaderno de Colab anexo, empleando regresiones logísticas, arboles de decisión y el modelo random forest, con el objetivo de predecir el tipo de movimiento según las coordenadas.

Al analizar la distribución de la variable objetivo Subtipo\_nombre, se observó que algunos subtipos presentaban únicamente un registro en todo el inventario. Este escenario impide el uso de una partición estratificada del conjunto de datos, ya que la clase menos representada debe contar con al menos dos observaciones para poder estar simultáneamente en los subconjuntos de entrenamiento y prueba.

Para garantizar una evaluación adecuada de los modelos, se decidió filtrar la base de datos y conservar únicamente aquellos subtipos con al menos dos registros. De esta forma, fue posible aplicar un `train_test_split` estratificado y obtener métricas de desempeño más estables y representativas.

Una vez se entrenaron y testearon los modelos encontramos los siguientes resultados, de los cuales inferimos:

## -Regresión logística

- Accuracy ≈ 0.35
- La mayoría de las clases tienen:
  - $precision = 0$
  - $recall = 0$
- El modelo básicamente se concentra en “deslizamiento translacional” y “deslizamiento rotacional”, que son las clases más frecuentes.
- En la matriz de confusión se ve que casi todas las filas descargan sus predicciones en esas dos clases.

La regresión logística no es capaz de separar bien tantos subtipos diferentes usando solo x,y con una frontera lineal. Tiende a colapsar muchas clases raras dentro de las más frecuentes.

## -Árbol de decisión

- Accuracy ≈ 0.36 (ligeramente mejor que la logística).
- Mejora algo las métricas de:
  - “deslizamiento rotacional”
  - “deslizamiento translacional”
  - “caída de detritos”, “caída de roca”, “reptación de suelos”, etc.
- Macro avg:
  - precision ~0.23
  - recall ~0.25
  - f1 ~0.23

Es decir, en promedio por clase, el desempeño es modesto, pero ya no es “todo cero” como en la logística.

E árbol sí captura relaciones no lineales en el espacio (x,y), pero sigue siendo un único modelo, limitado y con riesgo de sobreajuste local.

### -Random Forest

- Accuracy  $\approx 0.40$  → es el mejor de los tres.
- Se ve mejora en varias clases:
  - “deslizamiento rotacional”:  $f1 \approx 0.48$
  - “deslizamiento translacional”:  $f1 \approx 0.52$
  - “solifluxión”:  $f1 \approx 0.51$
  - “caída de detritos”, “caída de roca”, etc. mejoran un poco respecto al árbol.
- Las clases con muy pocos datos (1–5 registros) siguen con precision/recall muy bajos o 0.  
Eso es normal: no hay información suficiente para que el modelo aprenda patrones espaciales confiables.

E Random Forest es el modelo que mejor explota la relación espacial entre x,y y el subtipo, pero está limitado por:

- el fuerte desbalance de clases, y
- el hecho de que solo usamos coordenadas x,y (sin pendiente, litología, etc.).

¿Y los warnings de UndefinedMetricWarning?

Pasa porque hay clases para las cuales el modelo nunca predijo ningún ejemplo, entonces:

- precision = 0.0
- recall = 0.0
- f1 = 0.0

Es completamente esperable cuando:

- hay muchas clases,
- algunas son muy minoritarias,
- y el modelo se centra en las clases grandes.

Esto es una consecuencia del desbalance y la escasez de datos por clase.

## Comparación del rendimiento entre modelos

Para la fase de modelamiento se replanteó el problema de predicción con el fin de evitar la fuga de información identificada en los primeros experimentos. En lugar de predecir el tipo general de movimiento a partir de variables que duplicaban su contenido, se planteó como objetivo la clasificación del Subtipo\_nombre utilizando únicamente las coordenadas espaciales (x, y) de cada registro. De este modo, se evaluó hasta qué punto la ubicación geográfica se relaciona con el subtipo de movimiento en masa.

Se entrenaron tres modelos de clasificación supervisada sobre un conjunto de entrenamiento (70 %) y se evaluaron en un conjunto de prueba (30 %), manteniendo la proporción de clases mediante un train\_test\_split estratificado:

- Regresión Logística
- Árbol de Decisión
- Random Forest

En términos de desempeño global, las precisiones (accuracy) obtenidas fueron aproximadamente:

- Regresión logística: 0,35
- Árbol de decisión: 0,36
- Random Forest: 0,40

La regresión logística presentó las peores métricas por clase. Aunque su exactitud global alcanza el 35 %, el modelo tiende a concentrar la mayoría de las predicciones en los subtipos más frecuentes, como *deslizamiento rotacional* y *deslizamiento traslacional*, mientras que para la mayoría de los subtipos minoritarios la precisión y el recall son prácticamente nulos. Esto indica que una frontera de decisión lineal en el espacio definido por x e y no es suficiente para separar adecuadamente los distintos subtipos de movimiento.

El árbol de decisión mejora ligeramente el desempeño global (accuracy ≈ 0,36) y muestra valores de precision y recall moderados para algunos subtipos con mayor soporte, en particular los deslizamientos y ciertos tipos de caídas. Sin embargo, el promedio macro de las métricas (precision y recall alrededor de 0,23–0,25) sigue evidenciando dificultades para clasificar correctamente las clases menos representadas y una fuerte influencia del desbalance del conjunto de datos.

El modelo de Random Forest es el que presenta el mejor rendimiento general, con una exactitud cercana al 40 % y mejoras visibles en las métricas de varios subtipos relevantes. En este caso, subtipos como *deslizamiento rotacional*, *deslizamiento traslacional* y *solifluxión* alcanzan valores de f1-score moderados (del orden de 0,4–0,5), lo que sugiere que el ensamble de múltiples árboles permite capturar mejor patrones espaciales no lineales a partir de las coordenadas. No obstante, al igual que en el árbol de decisión, las clases con muy pocos ejemplos siguen presentando un desempeño deficiente, con precisión y recall cercanos a cero.

En síntesis, la comparación entre modelos muestra que: El uso exclusivo de las coordenadas x e y proporciona una capacidad de predicción limitada, aunque suficiente para obtener patrones aprovechables por modelos no lineales como Random Forest. Entre los tres algoritmos evaluados, Random Forest ofrece el mejor compromiso entre desempeño global y comportamiento por clase, por lo que puede considerarse el modelo más adecuado dentro de las restricciones del conjunto de variables disponibles.

## **Referencias**

**"SGC. 10 de marzo de 2023"** Inventario de movimientos en masa. Recuperado de:  
[https://datos.sgc.gov.co/datasets/312c8792ddb24954a9d2711bd89d1afe\\_0/explore?location=1.140926%2C-76.180844%2C6.56](https://datos.sgc.gov.co/datasets/312c8792ddb24954a9d2711bd89d1afe_0/explore?location=1.140926%2C-76.180844%2C6.56).

**"Guzzetti, Fausto. Et al. Abril de 2012."** Landslide inventory maps: New tools for an old problem. Recuperado de: <https://doi.org/10.1016/j.earscirev.2012.02.001>.