Proyecto Final

Introducción a la Gestión de Datos Geocientíficos

Juan Pablo Lozada Escobar

Docente: Ever Herrera Rios

Universidad Nacional de Colombia – Sede Medellín

Facultad de Minas

Diciembre de 2024

Contenido

01

Introducción

Contexto y propósito del estudio.

02

Generalidades del Dataset

Descripción técnica de las columnas del inventario.

03

Limpieza de los Datos

Proceso de depuración y estandarización.

04

Estadísticos del Conjunto de Datos

Análisis de variables categóricas y correlaciones.

05

Presentación General de los Modelos

Regresión Logística, Árbol de Decisión, Random Forest.

06

Comparación del Rendimiento

Evaluación y conclusiones sobre los modelos.

Introducción: Gestión de Datos Geocientíficos

Los inventarios de movimientos en masa son cruciales para entender la ocurrencia y caracterización de fenómenos como deslizamientos y flujos. La calidad de los datos es fundamental para desarrollar modelos confiables de análisis espacial y estimación de riesgo.

Este trabajo revisa el inventario del Servicio Geológico Colombiano, explorando sus variables y estructura. El objetivo es evaluar cómo el tratamiento, limpieza y modelamiento estadístico de este inventario puede mejorar la clasificación y predicción de tipos de movimientos en masa, y qué modelo ofrece el mejor desempeño.

Generalidades del Dataset Geoespacial

El dataset es un inventario de movimientos en masa y procesos gravitacionales, diseñado para un Sistema de Información Geográfica (SIG). Cada columna describe características geomorfológicas, administrativas o cartográficas de los eventos.

Identificadores Clave

- **IDENTIFICACIÓN DEL OBJETO:** Identificador geométrico único para trazabilidad e integridad.
- **IDENTIFICACIÓN:** ID numérico secuencial para referencia rápida y organización.
- **Inventario_Movimiento:** Valor cuantitativo asociado al inventario base para validación.

Clasificación de Movimientos

- **F35DOV_TIP:** Código de dominio estandarizado (ej., 67 = Caída, 69 = Deslizamiento).
- **Tipo_Movimiento:** Categoría principal del proceso (ej., Deslizamiento, Flujo, Caída).
- **Subtipo_Movimiento:** Código numérico del subtipo específico (ej., 1 = Volcamiento flexural de roca).
- **Subtipo_nombre:** Descripción literal del subtipo (ej., "Deslizamiento traslacional").
- **Etiqueta:** Rótulo simplificado para identificación rápida en mapas.

Justificación del Uso del Dataset en SIG

Este dataset es fundamental para el análisis, gestión y representación espacial de la inestabilidad del terreno. Su estructura permite una clasificación estandarizada y jerárquica de los movimientos, esencial para comprender la dinámica geomorfológica y desarrollar modelos de amenazas y riesgos.

Caracterización Precisa

Campos como **Tipo_Movimiento** y **Subtipo_Movimiento** permiten una caracterización detallada, alineada con clasificaciones técnicas aceptadas.

Estandarización Visual

Atributos como **Representación_mapa** y **Etiqueta** facilitan la consistencia visual en productos cartográficos oficiales.

Interoperabilidad y Trazabilidad

Identificadores únicos (**ESRI_OID** e **ID**) permiten relaciones robustas con otras bases de datos, asegurando la integridad del sistema.

Limpieza de los Datos

Para optimizar el dataset, se eliminaron columnas redundantes y se estandarizaron categorías. Las columnas **F35DOV_TIP** y **Subtipo_Movimiento** fueron eliminadas por ser redundantes con **Tipo_Movimiento** y **Subtipo_nombre** respectivamente. También se eliminó **Representacion_mapa** por contener información ya presente en **Subtipo_nombre**.

La estandarización de categorías fue crucial para corregir inconsistencias como mayúsculas/minúsculas mezcladas, tildes y espacios extras, lo que mejora el modelamiento.

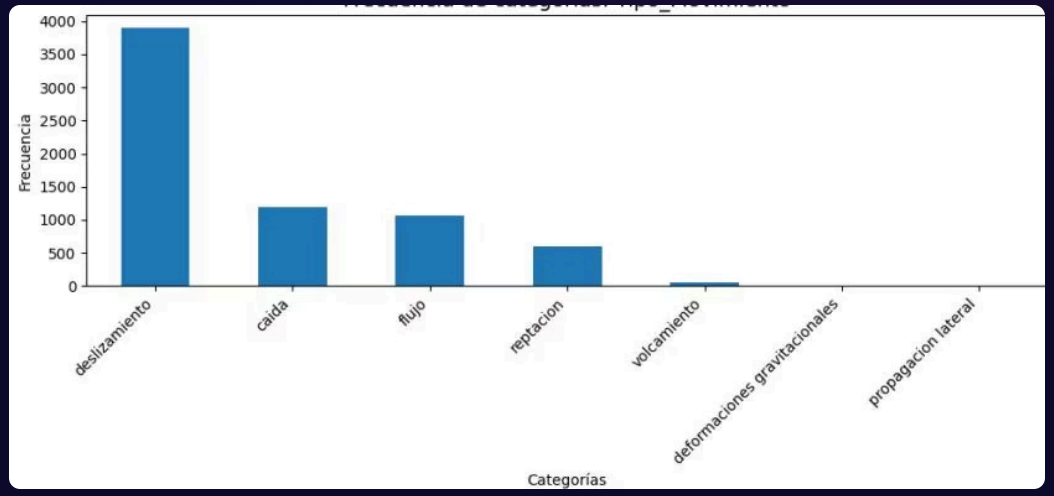
FID	int64	0	6826
OBJECTID	float64	0	4
Tipo_Movimiento	object	0	7
Subtipo_nombre	object	0	24
ESRI_OID	int64	0	6826

Estadísticos de Variables Categóricas

Se analizaron las frecuencias absolutas y relativas de las variables categóricas para entender la distribución de los tipos y subtipos de movimiento. "Deslizamiento" y "deslizamiento traslacional" son las categorías más representativas.

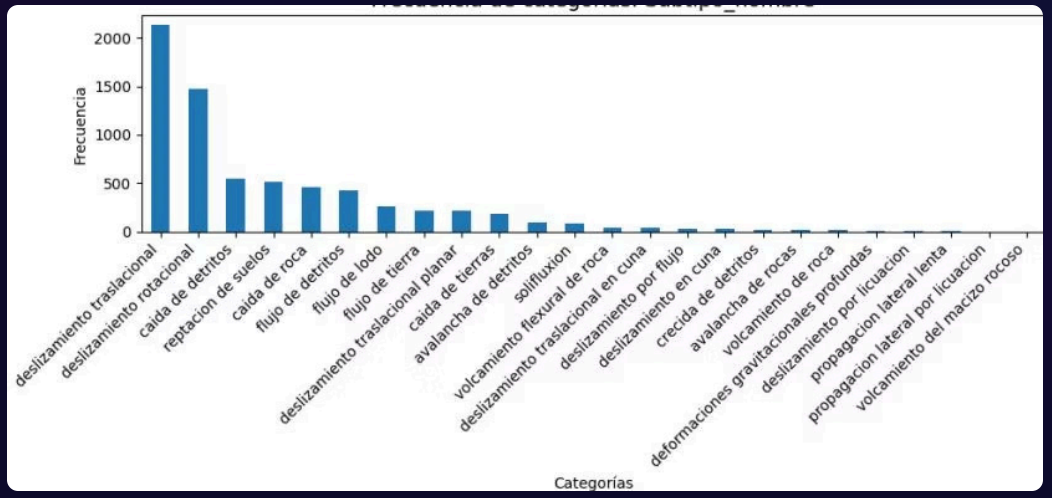
Tipo_Movimiento

- Deslizamiento: 3900 (57.1%)
- Caída: 1196 (17.5%)
- Flujo: 1064 (15.6%)
- Reptación: 595 (8.7%)



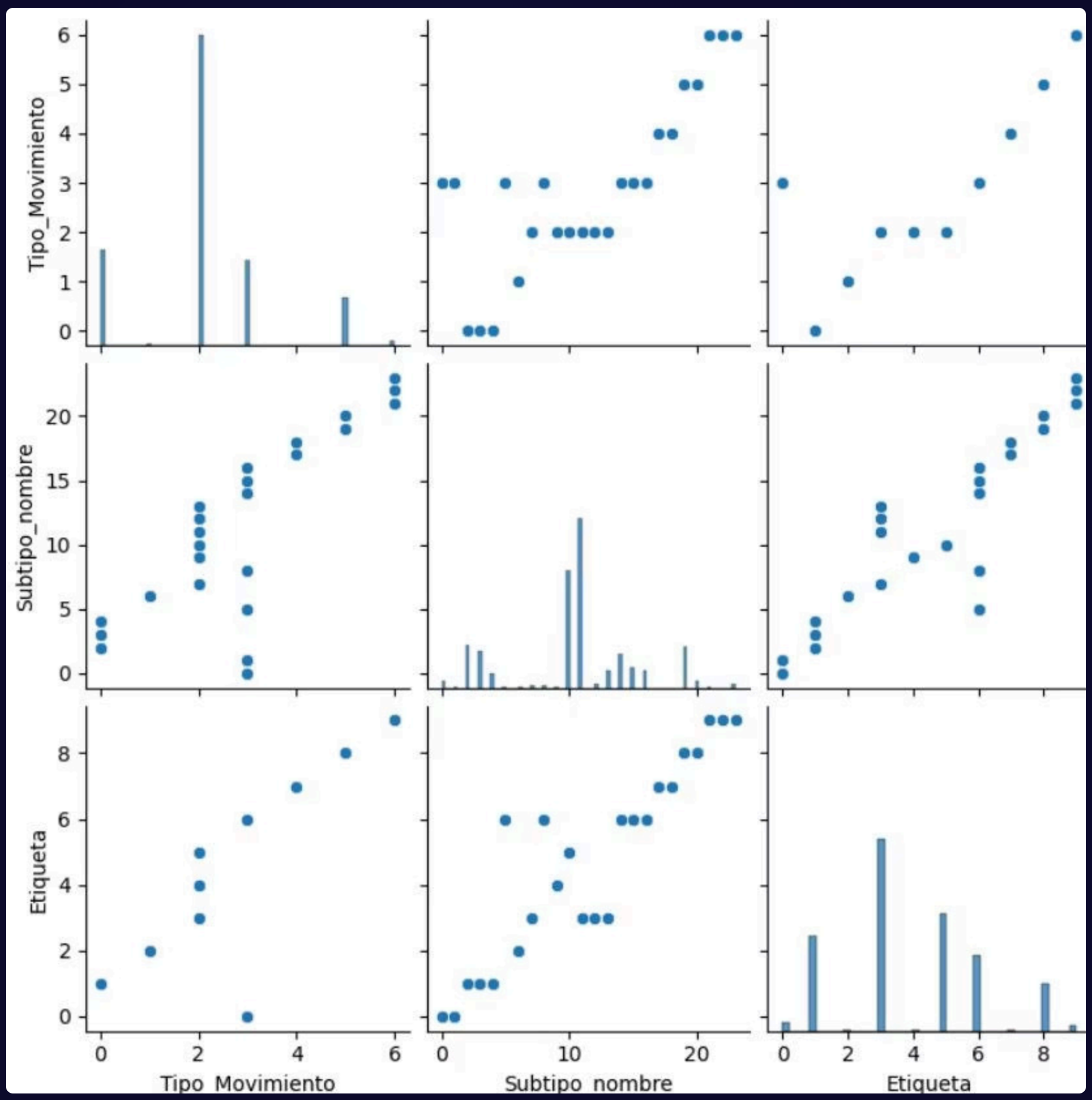
Subtipo_nombre

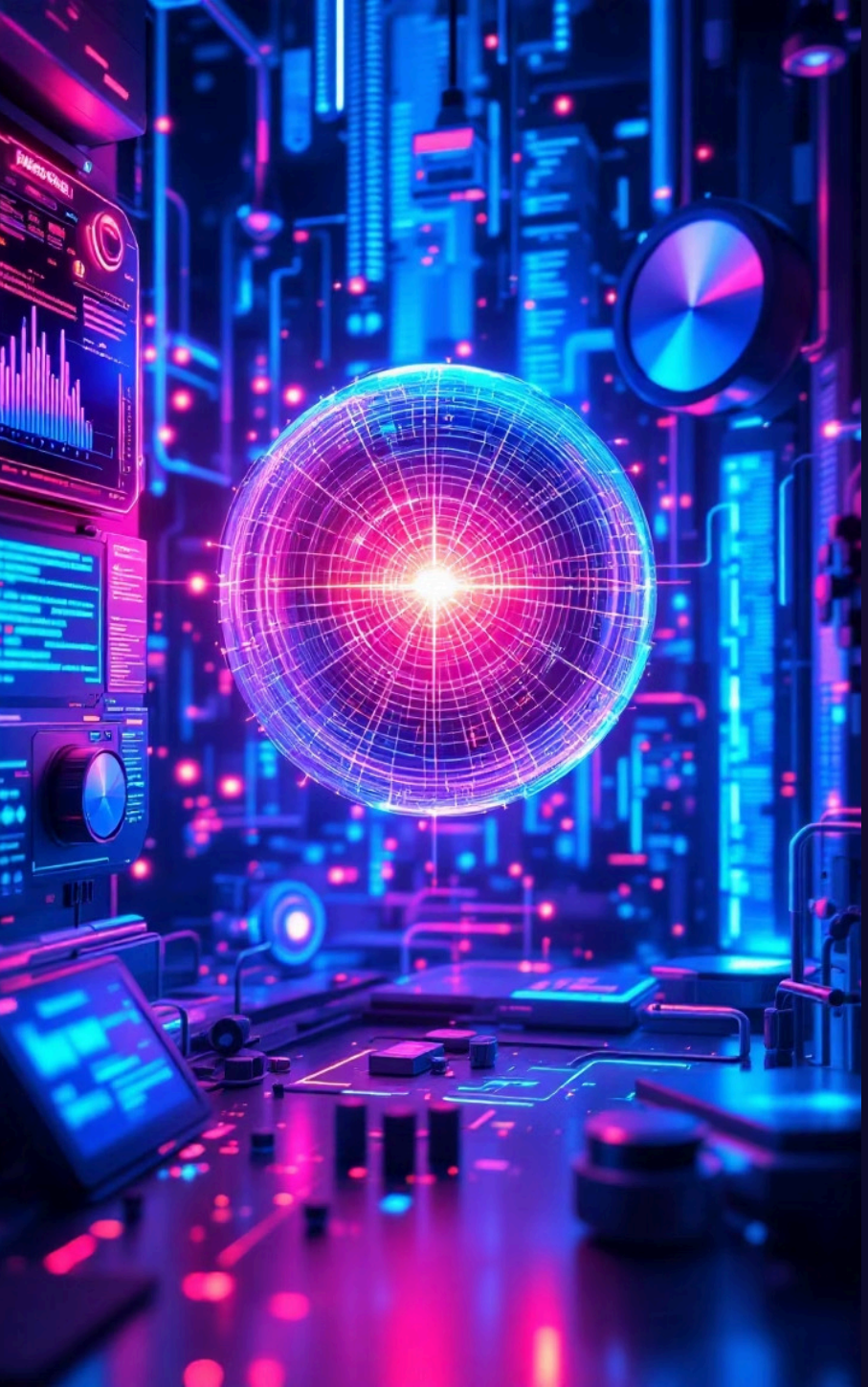
- Deslizamiento traslacional: 2130 (31.2%)
- Deslizamiento rotacional: 1475 (21.6%)
- Caída de detritos: 547 (8.0%)
- Reptación de suelos: 515 (7.5%)



Correlación entre Variables Categóricas

Se generó un heatmap para explorar patrones de co-ocurrencia entre las variables temáticas. Existe una fuerte correlación entre **Etiqueta**, **Subtipo_nombre** y **Tipo_Movimiento**, lo que indica una relación consistente entre estas clasificaciones.





Presentación General de los Modelos

Se entrenaron modelos de Regresión Logística, Árbol de Decisión y Random Forest para predecir el **Subtipo_nombre** usando solo coordenadas (x,y). Se filtraron subtipos con menos de dos registros para permitir una partición estratificada del dataset.

1

Regresión Logística

Accuracy ≈ 0.35. Baja precisión y recall para la mayoría de las clases, concentrándose en "deslizamiento traslacional" y "deslizamiento rotacional".

2

Árbol de Decisión

Accuracy ≈ 0.36. Mejora ligeramente las métricas para subtipos más frecuentes. Macro avg precision y recall ~0.23-0.25.

3

Random Forest

Accuracy ≈ 0.40. El mejor de los tres, con mejoras en varios subtipos como "deslizamiento rotacional" y "solifluxión".

Comparación del Rendimiento entre Modelos

El problema de clasificación es complejo debido al número de subtipos y al desbalance de clases. El uso exclusivo de coordenadas (x, y) limita la predicción, pero modelos no lineales como Random Forest logran capturar patrones espaciales.

Regresión Logística

Peor desempeño, con una exactitud del 35%. Concentra predicciones en subtipos frecuentes, con métricas nulas para minoritarios.

Árbol de Decisión

Mejora ligeramente con un 36% de exactitud. Desempeño modesto para clases menos representadas.

Random Forest

Mejor rendimiento general, con un 40% de exactitud. Captura mejor patrones espaciales no lineales, aunque clases minoritarias siguen siendo un desafío.

Random Forest ofrece el mejor compromiso entre desempeño global y comportamiento por clase, siendo el modelo más adecuado dadas las restricciones de variables.