Exploración y Preprocesamiento del Dataset

```
Comienza a programar o generar con IA.
```

1. Introducción

Entrega 2 del proyecto de Aprendizaje Automático."

Objetivo: construir un dataset unificado para predecir el riesgo de inundación en barrios populares del Gran Buenos Aires.

Vamos a unir cuatro datasets geoespaciales, construir variables predictoras, y preparar el dataset para modelado supervisado.

2. Preparación del entorno Carga de Datasets

```
# Carga manual de archivos
from google.colab import files
uploaded = files.upload()
```

```
Elegir archivos 4 archivos
du_techo_cuerpos_de_agua.geojson(n/a) - 312782 bytes, last modified: 2/10/2025 - 100% done
du_techo_cursos_de_agua.geojson(n/a) - 1850932 bytes, last modified: 2/10/2025 - 100% done
du_techo_microbasurales.geojson(n/a) - 1178997 bytes, last modified: 2/10/2025 - 100% done
du_techo_zonas_inundables.csv(text/csv) - 1491171 bytes, last modified: 2/10/2025 - 100% done
Saving du_techo_cuerpos_de_agua.geojson to du_techo_cuerpos_de_agua.geojson
Saving du_techo_cursos_de_agua.geojson to du_techo_cursos_de_agua.geojson
Saving du_techo_microbasurales.geojson to du_techo_microbasurales.geojson
Saving du_techo_zonas_inundables.csv to du_techo_zonas_inundables.csv
```

```
# Preparamos el entorno
import pandas as pd
import geopandas as gpd

# Cargar los datasets
zonas_inundables = pd.read_csv("du_techo_zonas_inundables.csv")
cuerpos_agua = gpd.read_file("du_techo_cuerpos_de_agua.geojson")
cursos_agua = gpd.read_file("du_techo_cursos_de_agua.geojson")
microbasurales = gpd.read_file("du_techo_microbasurales.geojson")
```

En esta sección se cargan los cuatro datasets que componen la base de datos del proyecto.

du_techo_zonas_inundables.csv contiene información tabular sobre zonas afectadas por inundaciones, incluyendo la variable objetivo se_inunda_.

du_techo_cuerpos_de_agua.geojson, du_techo_cursos_de_agua.geojson y du_techo_microbasurales.geojson son capas geoespaciales en formato GeoJSON que representan elementos ambientales relevantes. Se utilizan pandas para el archivo CSV y geopandas para los archivos GeoJSON.

Todos los archivos fueron obtenidos de fuentes públicas: DU-Techo y RENABAP (01//10/2025), y se subieron manualmente a Colab para su procesamiento. Dataset (fuente)

https://www.datos.gob.ar/dataset/habitat-factores-riesgo-barrios-populares-gran-buenos-aires

3. Exploración Inicial y Codificación de la Variable Objetivo

Vamos a analizar como está estructurado el dataset principal (du_techo_zonas_inundables.csv) y preparar la variable se_inunda_ para el modelo.

```
# Exploramos la estructura del dataset
zonas_inundables.info() # vemos estructura general
zonas_inundables.head() # vemos primeras filas
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2746 entries, 0 to 2745
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   WKT         2746 non-null   object 
 1   id          2746 non-null   int64  
 2   id_poligon  2741 non-null   float64
 3   se_inunda_  2746 non-null   object 
 4   con_que_fr  2746 non-null   object 
 5   provinicia  2746 non-null   object 
 6   departamen  2746 non-null   object 
 7   localidad   2746 non-null   object 
dtypes: float64(1), int64(1), object(6)
memory usage: 171.8+ KB
```

	WKT	id	id_poligon	se_inunda_	con_que_fr	provinicia	depar
0	MULTIPOLYGON (((-58.2143471291028 -26.21995308...	1	17325.0	SÍ, SÓLO EN UN SECTOR DEL BARRIO	SOLAMENTE CUANDO DILUVIA (UNA O DOS VECES POR ...	Formosa	For
1	MULTIPOLYGON (((-58.2179226698168 -26.20925840...	2	17322.0	SÍ, SÓLO EN UN SECTOR DEL BARRIO	CADA VEZ QUE LLUEVE FUERTE (MUCHAS VECES POR AÑO)	Formosa	For
2	MULTIPOLYGON (((-58.2334597447517 -26.20322391...	3	18578.0	SÍ, TODO EL BARRIO	SOLAMENTE CUANDO DILUVIA (UNA O DOS VECES POR ...	Formosa	For
3	MULTIPOLYGON (((-58.2764032110572 -26.27741542...	4	12001.0	SÍ, SÓLO EN UN SECTOR DEL BARRIO	SOLAMENTE CUANDO DILUVIA (UNA O DOS VECES POR ...	Formosa	For
4	MULTIPOLYGON (((-58.2777912542224 -26.27851061...	5	12000.0	SÍ, SÓLO EN UN SECTOR DEL BARRIO	SOLAMENTE CUANDO DILUVIA (UNA O DOS VECES POR ...	Formosa	For

Próximos pasos: Generar código con `zonas_inundables` New interactive sheet

```
# Vemos distribución de la variable objetivo
# (incluyendo si hay celdas con valores nulos)
```

```
zonas_inundables['se_inunda_'].value_counts(dropna=False)
```

se_inunda_	count
SÍ, TODO EL BARRIO	1374
SÍ, SÓLO EN UN SECTOR DEL BARRIO	1372

dtype: int64

Se utiliza value_counts(dropna=False) para contar todas las categorías presentes en la variable se_inunda_, incluyendo los valores faltantes (NaN). Esto permite detectar si hay registros incompletos que podrían afectar el modelado o requerir imputación o descarte.

```
# Inspeccionamos el nombre de todas las columnas del dataset
zonas_inundables.columns
```

```
Index(['WKT', 'id', 'id_poligon', 'se_inunda_', 'con_que_fr', 'provinicia',
       'departamen', 'localidad'],
      dtype='object')
```

Se inspeccionan las columnas del dataset zonas_inundables para identificar las variables disponibles.

Las columnas incluyen:

identificadores (id, id poligon),

información geográfica (provincia, departamen, localidad), y

variables relacionadas con el riesgo de inundación (se_inunda_, con_que_fr).

la columna WKT, contiene geometrías en formato texto

```
# Exploramos los valores únicos de las columnas categóricas
# Imprime los valores únicos
for col in zonas_inundables.columns:
    if zonas_inundables[col].dtype == 'object':
        print(f"\nValores únicos en {col}:")
        print(zonas_inundables[col].unique())
```

```
'El Alcázar' 'Puerto Rico' 'San Salvador De Jujuy' 'Monterrico' 'Palpalá'
'San Pedro De Jujuy' 'Humahuaca' 'Libertador General San Martín'
'Los Lapachos' 'Aguas Calientes' 'Maimará' 'La Quiaca' 'Caimancito'
'Rawson' 'Trelew' 'Puerto Madryn' 'Comodoro Rivadavia' 'Concepción'
'Lago Puelo' 'Neuquén' 'Centenario' 'Plottier' 'Junín De Los Andes'
'Aluminé' 'Senillosa' 'San Miguel De Tucumán' 'Ingenio La Florida'
'Alderetes' 'Banda Del Río Salí' 'Aguilares' 'Famaillá'
'Juan Bautista Alberdi' 'San Isidro De Lules' 'Acheral' 'El Manantial'
'Yerba Buena - Marcos Paz' 'Ingenio San Pablo' 'Villa Carmela'
'Las Talitas' 'Alto Verde' 'Arcadia' 'San Roque' 'La Reducción'
'Diagonal Norte, Luz Y Fuerza, Los Pocitos, Villa Nueva Italia'
'Tafí Viejo' 'Barrio San Felipe' 'Río Chico' 'Garmendia'
'Villa Burruyacú' 'Ex Ingenio Los Ralos' 'Río Colorado' 'Santa Lucía'
'Santa Ana' 'Simoca' 'Amaicha Del Valle' 'Tafí Del Valle'
'Villa De Trancas' 'Córdoba Capital' 'Estacion Juárez Celman'
'Villa Allende' 'Alta Gracia' 'Oncativo' 'Unquillo' 'Jesús María'
'Morteros' 'Cosquín' 'Villa Carlos Paz' 'Río Cuarto' 'Bouwer'
'Río Tercero' 'Bell Ville' 'Villa Nueva' 'La Calera' 'Montecristo'
'Cruz Del Eje' 'Villa Giardino' 'Río Segundo' 'Santa Ana De Los Guácaras'
'Corrientes' 'Laguna Brava' 'Saladas' 'Goya' 'Esquina'
'San Luis Del Palmar' 'Paso De Los Libres' 'Concordia' 'Santo Tomé'
'Empedrado' 'Mburucuyá' 'Monte Caseros' 'Paraná' 'Colonia Avellaneda'
'Concepción Del Uruguay' 'Gualeguay' 'Gualeguaychú' 'Nogoyá' 'Victoria'
'Rosario Del Tala' 'Federación' 'Colón' 'La Paz' 'Rincón de Nogoyá'
'Diamante' 'Villaguay' 'Los Corralitos' 'Las Heras' 'Panquehua'
'El Borbollón' 'Mendoza' 'Guaymallén' 'Fray Luis Beltrán' 'Russell'
'Palmira' 'Villa Tulumaya' 'Rodeo Del Medio' 'Tunuyán' 'General Alvear'
'El Plumerillo' 'Sarmiento' 'Mayor Drummond' 'Trapiche' 'Buena Nueva'
'Bermejo' 'El Challao' 'Chacras De Coria' 'Jesús Nazareno'
'Rodeo De La Cruz' 'Perdriel' 'Colonia Segovia'
'Vertientes del Pedemonte' 'San Martín' 'Rivadavia' 'La Arboleda'
'Malargüe' 'San Rafael' 'El Peral' 'Cuadro Benegas' 'Las Paredes' 'Maipú'
'Eugenio Bustos' 'Colonia Molina' 'El Sauce' 'Ugarteche' 'Agrelo'
'Uspallata' 'Vista Flores' 'Puente De Hierro' 'Goudge' 'Cipolletti'
'Allen' 'General Roca' 'Cinco Saltos' 'Balsa Las Perlas'
'San Carlos De Bariloche' 'El Bolsón' 'Viedma' 'Villa Regina'
'San Antonio Oeste' 'Mainqué' 'Las Grutas' 'Cervantes' 'Villa Mercedes'
'San Luis' 'Juana Koslay' 'Rosario' 'Villa Gobernador Gálvez' 'Pérez'
'Granadero Baigorria' 'Alvear' 'Capitán Bermúdez' 'Roldán' 'San Lorenzo'
'Pueblo Esther' 'General Lagos' 'Arroyo Seco' 'Sauce Viejo'
'Santa Rosa De Calchines' 'San José Del Rincón' 'Santa Fe' 'Arroyo Leyes'
'Venado Tuerto' 'Calchaquí' 'Casilda' 'Coronda' 'Esperanza' 'Frontera'
'Laguna Paiva' 'Las Toscas' 'Rafaela' 'San Cristóbal' 'San Jorge'
'Tostado' 'Villa Constitución' 'Villa Ocampo' 'Reconquista' 'Correa'
'Médano De Oro' 'Totoras' 'Tacuarendí' 'Piñero' 'Barrancas'
'Cañada De Gómez' 'Oliveros' 'Cayastá' 'Helvecia' 'Ingeniero Chanourdie'
'Chimbas' 'La Bebida' 'Caucete' 'Pie De Palo' 'Las Talas' 'San Juan'
'Ullum' 'Villa Borjas' 'Villa Santa Rosa' 'Vallecito' 'Tres Esquinas'
'Villa Media Agua' 'Campo de Herrera' 'Firmat']
```

Interpretación de la exploración inicial

Variable se_inunda_

Valores únicos encontrados:

"SÍ, TODO EL BARRIO" (1374 registros)

"SÍ, SÓLO EN UN SECTOR DEL BARRIO" (1372 registros)

No se detectaron registros con la categoría "NO SE INUNDA", lo cual indica que el dataset está compuesto exclusivamente por zonas que presentan algún nivel de inundación. Es por ello que tomamos la decisión de trabajar con una clasificación binaria supervisada, diferenciando entre inundación total (2) e inundación parcial (1).

Variable con_que_fr

Valores únicos encontrados:

"CADA VEZ QUE LLUEVE FUERTE (MUCHAS VECES POR AÑO)" (1488)

"OCASIONALMENTE (ALGUNAS VECES POR AÑO)" (638)

"SOLAMENTE CUANDO DILUVIA (UNA O DOS VECES POR AÑO)" (572)

"Otro, especificar" (43)

"NS/NC" (2)

Otros valores con muy baja frecuencia (1–2 registros)

Esta variable aporta información sobre la frecuencia de inundación, lo cual puede ser útil como variable predictora. Se observa una predominancia de zonas que se inundan con frecuencia alta. Podríamos agrupar en categorías (Alta, Media, Baja) para facilitar el modelado.

Variable provincia, departamento y localidad

Valores únicos encontrados: Más de 20 provincias. Más de 50 departamentos. Más de 100 localidades. En los 3 casos, aunque las variables aportan contexto geográfico, se decide no incluirlas como variable predictora debido a que:

Tienen muchas categorías.

La ubicación ya está representada por coordenadas (latitud, longitud).

Podría introducir ruido en el modelo y sobreajuste.

Variable WKT

Valores únicos: Geometrías en formato MULTIPOLYGON

Esta columna contiene la geometría de cada zona en formato texto. No se usará directamente en el modelo, pero se puede utilizar para calcular variables derivadas como el centroide o el área.

```
# Vemos la distribución de la variable con_que_fr
zonas_inundables['con_que_fr'].value_counts(dropna=False)
```

con_que_fr	count
CADA VEZ QUE LLUEVE FUERTE (MUCHAS VECES POR AÑO)	1488
OCASIONALMENTE (ALGUNAS VECES POR AÑO)	638
SOLAMENTE CUANDO DILUVIA (UNA O DOS VECES POR AÑO)	572
Otro, especificar	43
NS/NC	2
CADA VEZ QUE LLUEVE (MUCHAS VECES POR AÑO)	2
CADA VEZ QUE LLUEVE FUERTE (MUCHAS VECES AL AÑO)	1

dtype: int64

Interpretación

La mayoría de las zonas se inundan frecuentemente: más de la mitad (1488 de 2746) reportan que se inundan cada vez que llueve fuerte.

Hay una proporción significativa que se inunda ocasionalmente (638) o solo en eventos extremos (572).

Los valores "Otro, especificar" y "NS/NC" son pocos, pero podrían representar respuestas abiertas o falta de información.

Aparecen duplicados con pequeñas variaciones en la redacción, como "CADA VEZ QUE LLUEVE FUERTE (MUCHAS VECES POR AÑO)" y "CADA VEZ QUE LLUEVE FUERTE (MUCHAS VECES AL AÑO)", vamos a normalizar estos textos.

```
# Preprocesamiento (agrupar en tres categorías simplificadas)
# Transformación de variable categórica textual (con_que_fr) en una versión más
def codificar_frecuencia(valor):
    valor = str(valor).upper()
    if "CADA VEZ" in valor or "MUCHAS VECES" in valor:
        return "Alta"
    elif "OCASIONALMENTE" in valor:
        return "Media"
    elif "SOLAMENTE" in valor or "UNA O DOS VECES" in valor:
        return "Baja"
    else:
        return "Otro"

zonas_inundables['frecuencia_codificada'] = zonas_inundables['con_que_fr'].apply

zonas_inundables['frecuencia_codificada'].value_counts(dropna=False)
```

frecuencia_codificada	count
Alta	1491
Media	638
Baja	572
Otro	45

dtype: int64

Se agrupan los valores en cuatro categorías:

Alta: zonas que se inundan cada vez que llueve fuerte o muchas veces por año.

Media: zonas que se inundan ocasionalmente.

Baja: zonas que solo se inundan en eventos extremos.

Otro: respuestas abiertas o no clasificables.

Esta transformación permite reducir la variabilidad textual y facilita el uso de la variable como predictora en el modelo supervisado.

Resultados de la exploración:

Tras analizar las columnas se_inunda_ y con_que_fr, se observa que el dataset no contiene registros con la categoría "NO SE INUNDA". Esto indica que el archivo du_techo_zonas_inundables.csv está compuesto exclusivamente por zonas afectadas por inundaciones, ya sea total o parcialmente. Por lo tanto, se ajusta la variable objetivo a una clasificación binaria:

1: Inundación parcial

2: Inundación total

Esta decisión permite construir un modelo supervisado que prediga el grado de afectación en zonas vulnerables, cumpliendo con el objetivo del proyecto.

Además, se analiza la variable con_que_fr, que describe la frecuencia con la que se inunda cada zona. Para facilitar su uso como variable predictora, se realiza un preprocesamiento que agrupa sus valores textuales en cuatro categorías simplificadas:

Alta: Zonas que se inundan cada vez que llueve fuerte o muchas veces por año

Media: Zonas que se inundan ocasionalmente

Baja: Zonas que solo se inundan en eventos extremos