

# Proyecto\_Fase1

Laura\_Marlon\_Andrés\_Eduar

2025-09-03

## R Markdown

### 1. Definición del problema

#### Descripción clara del problema

El proyecto busca analizar el riesgo de inundaciones pluviales urbanas en segmentos micro-urbanos de ciudades globales, utilizando el dataset sintético `urban_pluvial_flood_risk_dataset.csv`. Este dataset proporciona información sobre factores como elevación, densidad de drenaje, uso del suelo, intensidad de lluvia histórica y etiquetas de riesgo. El objetivo es identificar patrones y correlaciones que permitan entender la vulnerabilidad a inundaciones y apoyar estrategias de mitigación y planificación urbana.

#### Justificación de su importancia

Las inundaciones pluviales urbanas son un desafío global creciente debido al cambio climático, la urbanización rápida y la infraestructura de drenaje insuficiente. Según el Banco Mundial, las inundaciones afectan a más de 20 millones de personas anualmente, con pérdidas económicas que superan los \$40 mil millones. Este análisis es crucial para desarrollar políticas de resiliencia urbana, especialmente en ciudades de países en desarrollo y desarrollados con alta densidad urbana. La minería de datos permite identificar áreas críticas y predecir riesgos, optimizando recursos para prevención.

#### Preguntas de investigación o hipótesis

- **Preguntas:**

- ¿Qué ciudades presentan mayor frecuencia de etiquetas de riesgo como `ponding_hotspot` o `extreme_rain_history`?
- ¿Cómo influyen el uso del suelo (`land_use`) y el grupo de suelo (`soil_group`) en el riesgo de inundaciones?
- ¿Qué variables numéricas (elevación, intensidad de lluvia, densidad de drenaje) tienen mayor correlación con el riesgo?

- **Hipótesis:**

- H1: Segmentos con baja elevación (`elevation_m < 10`) y baja densidad de drenaje (`drainage_density_km_per_km2 < 5`) tienen mayor probabilidad de etiquetas como `low_lying`.
- H2: Áreas con uso del suelo `Residential` o `Commercial` y grupo de suelo D (baja infiltración) presentan mayor riesgo de `ponding_hotspot`.

## 2. Descripción del dataset

### Nombre y fuente del dataset

- **Nombre:** Urban Pluvial Flood Risk Dataset (Global City Analysis 2025)
- **Fuente:** Kaggle, subido por Pratyush Puri (enlace)
- **Archivo principal:** `urban_pluvial_flood_risk_dataset.csv`
- **Detalles:** Dataset sintético con ~2963 registros de segmentos urbanos en 58 ciudades globales, enfocado en factores de riesgo de inundaciones pluviales.

### Variables disponibles y su significado

El dataset contiene 17 variables: 1. **segment\_id** (categórica): Identificador único del segmento. 2. **city\_name** (categórica): Nombre de la ciudad (ej. “Colombo, Sri Lanka”). 3. **admin\_ward** (categórica): Distrito o barrio administrativo. 4. **latitude** (numérica): Latitud del segmento. 5. **longitude** (numérica): Longitud del segmento. 6. **catchment\_id** (categórica): ID de la cuenca hidrológica. 7. **elevation\_m** (numérica): Elevación en metros, afecta el drenaje. 8. **dem\_source** (categórica): Fuente del modelo digital de elevación (ej. Copernicus\_GLO-30\_v2023). 9. **land\_use** (categórica): Uso del suelo (ej. Residential, Industrial, Roads, Green). 10. **soil\_group** (categórica): Grupo de suelo basado en infiltración (A: alta, D: baja). 11. **drainage\_density\_km\_per\_km2** (numérica): Densidad de drenaje en km/km². 12. **storm\_drain\_proximity\_m** (numérica): Proximidad al drenaje de tormenta en metros. 13. **storm\_drain\_type** (categórica): Tipo de drenaje (ej. CurbInlet, OpenChannel, None). 14. **rainfall\_source** (categórica): Fuente de datos de precipitación (ej. ERA5, IMD). 15. **historical\_rainfall\_intensity\_mm\_hr** (numérica): Intensidad histórica de lluvia en mm/h. 16. **return\_period\_years** (numérica): Período de retorno de eventos de lluvia (ej. 5, 25, 100). 17. **risk\_labels** (categórica): Etiquetas de riesgo (múltiple, ej. monitor, ponding\_hotspot, low\_lying).

### Identificación de posibles problemas iniciales

- **Valores faltantes:** Columnas como `elevation_m`, `soil_group`, `drainage_density_km_per_km2`, `storm_drain_proximity_m`, `storm_drain_type`, y `rainfall_source` presentan valores faltantes.
- **Ruido:** Elevaciones negativas (ej. -3 m) pueden indicar áreas bajo el nivel del mar o errores sintéticos.
- **Otros:** Posibles duplicados en `segment_id`, desbalance en ciudades (algunas con más segmentos), multicolinealidad (ej. `elevation_m` y `risk_labels` como `low_lying`), y etiquetas múltiples en `risk_labels` que requieren procesamiento.

## 3. Análisis Exploratorio de Datos (EDA) Básico

### 3.1 Carga y visualización general de los datos

```
## segment_id      city_name  admin_ward  latitude longitude
## 1 SEG-00001    Colombo, Sri Lanka Borough East   6.920633  79.91260
## 2 SEG-00002      Chennai, India      Ward D   13.076487  80.28177
## 3 SEG-00003    Ahmedabad, India    Sector 12  23.019473  72.63858
## 4 SEG-00004    Hong Kong, China    Sector 14  22.302602  114.07867
## 5 SEG-00005 Durban, South Africa    Sector 5 -29.887602  30.91101
## catchment_id elevation_m      dem_source      land_use soil_group
## 1      CAT-136          NA Copernicus_EEA-10_v5 Institutional
## 2      CAT-049       -2.19 Copernicus_EEA-10_v5  Residential      D
## 3      CAT-023       30.88          SRTM_3arc    Industrial      B
## 4      CAT-168       24.28          SRTM_3arc    Residential      B
## 5      CAT-171       35.70          SRTM_3arc    Industrial      C
```

```
## drainage_density_km_per_km2 storm_drain_proximity_m storm_drain_type
## 1 4.27 160.5 CurbInlet
## 2 7.54 NA OpenChannel
## 3 11.00 152.5 OpenChannel
## 4 7.32 37.0 Manhole
## 5 4.50 292.4 OpenChannel
## rainfall_source historical_rainfall_intensity_mm_hr return_period_years
## 1 ERA5 39.4 50
## 2 ERA5 56.8 25
## 3 IMD 16.3 5
## 4 ERA5 77.0 10
## 5 ERA5 20.8 5
## risk_labels
## 1 monitor
## 2 ponding_hotspot|low_lying|event_2025-05-02
## 3 monitor
## 4 monitor
## 5 monitor

## Filas: 2963
## Columnas: 17
```

### 3.2 Estadísticas básicas

```
## 'data.frame': 2963 obs. of 17 variables:
## $ segment_id : chr "SEG-00001" "SEG-00002" "SEG-00003" "SEG-00004" ...
## $ city_name : chr "Colombo, Sri Lanka" "Chennai, India" "Ahmedabad, India" ...
## $ admin_ward : chr "Borough East" "Ward D" "Sector 12" "Sector 14" ...
## $ latitude : num 6.92 13.08 23.02 22.3 -29.89 ...
## $ longitude : num 79.9 80.3 72.6 114.1 30.9 ...
## $ catchment_id : chr "CAT-136" "CAT-049" "CAT-023" "CAT-168" ...
## $ elevation_m : num NA -2.19 30.88 24.28 35.7 ...
## $ dem_source : chr "Copernicus_EEA-10_v5" "Copernicus_EEA-10_v5" "SRTM_3ar" ...
## $ land_use : chr "Institutional" "Residential" "Industrial" "Residential" ...
## $ soil_group : chr "" "D" "B" "B" ...
## $ drainage_density_km_per_km2 : num 4.27 7.54 11 7.32 4.5 8.97 8.25 5.88 7.79 NA ...
## $ storm_drain_proximity_m : num 160 NA 152 37 292 ...
## $ storm_drain_type : chr "CurbInlet" "OpenChannel" "OpenChannel" "Manhole" ...
## $ rainfall_source : chr "ERA5" "ERA5" "IMD" "ERA5" ...
## $ historical_rainfall_intensity_mm_hr: num 39.4 56.8 16.3 77 20.8 ...
## $ return_period_years : int 50 25 5 10 5 50 10 10 10 5 ...
## $ risk_labels : chr "monitor" "ponding_hotspot|low_lying|event_2025-05-02" ...

## elevation_m drainage_density_km_per_km2 storm_drain_proximity_m
## Min. : -3.000 Min. : 1.270 Min. : 0.20
## 1st Qu.: 8.725 1st Qu.: 4.670 1st Qu.: 47.98
## Median : 25.130 Median : 6.250 Median : 91.70
## Mean : 37.690 Mean : 6.291 Mean : 123.20
## 3rd Qu.: 59.620 3rd Qu.: 7.830 3rd Qu.: 162.62
## Max. : 266.700 Max. : 12.070 Max. : 751.70
## NA's : 161 NA's : 284 NA's : 239
## historical_rainfall_intensity_mm_hr return_period_years
## Min. : 5.40 Min. : 2.00
```

```
## 1st Qu.: 25.80      1st Qu.: 5.00
## Median : 37.90      Median : 10.00
## Mean   : 43.81      Mean   : 19.73
## 3rd Qu.: 55.55      3rd Qu.: 25.00
## Max.    :150.00      Max.    :100.00
##
```

```
## Elevation_m: Media = 37.68982  Mediana = 25.13  SD = 38.70896
```

```
##
## Commercial      Green      Industrial      Informal Institutional
##      493      359      357      29      106
##      Mixed      Residential      Roads      Water
##      110      827      599      83
```

```
##
## Commercial      Green      Industrial      Informal Institutional
## 16.6385420      12.1160985      12.0485994      0.9787378      3.5774553
##      Mixed      Residential      Roads      Water
## 3.7124536      27.9109011      20.2159973      2.8012150
```

```
##
##      A      B      C      D
## 362 547 747 713 594
```

```
##
##      A      B      C      D
## 12.21735 18.46102 25.21093 24.06345 20.04725
```

```
##
## Accra, Ghana      Ahmedabad, India      Amsterdam, Netherlands
##      45      52      46
## Athens, Greece      Auckland, New Zealand      Bangkok, Thailand
##      58      48      35
## Barcelona, Spain      Bengaluru, India      Bogotá, Colombia
##      54      55      42
## Brisbane, Australia      Buenos Aires, Argentina      Cape Town, South Africa
##      48      50      50
## Chennai, India      Colombo, Sri Lanka      Copenhagen, Denmark
##      50      30      50
## Delhi, India      Dhaka, Bangladesh      Doha, Qatar
##      56      47      37
## Dubai, UAE      Durban, South Africa      Guangzhou, China
##      43      54      38
## Hamburg, Germany      Hanoi, Vietnam      Ho Chi Minh City, Vietnam
##      37      40      49
## Hong Kong, China      Houston, USA      Hyderabad, India
##      34      51      39
## Istanbul, Türkiye      Jakarta, Indonesia      Karachi, Pakistan
##      55      39      47
## Kolkata, India      Kuala Lumpur, Malaysia      Lagos, Nigeria
##      43      42      51
```

|    |                      |                   |                        |
|----|----------------------|-------------------|------------------------|
| ## | Lima, Peru           | London, UK        | Manila, Philippines    |
| ## | 55                   | 29                | 61                     |
| ## | Mexico City, Mexico  | Miami, USA        | Montreal, Canada       |
| ## | 57                   | 46                | 48                     |
| ## | Mumbai, India        | Nairobi, Kenya    | New Orleans, USA       |
| ## | 39                   | 47                | 50                     |
| ## | New York, USA        | Osaka, Japan      | Paris, France          |
| ## | 43                   | 38                | 52                     |
| ## | Philadelphia, USA    | Pune, India       | Rio de Janeiro, Brazil |
| ## | 59                   | 41                | 50                     |
| ## | Riyadh, Saudi Arabia | Rome, Italy       | Rotterdam, Netherlands |
| ## | 46                   | 51                | 58                     |
| ## | San Francisco, USA   | Sao Paulo, Brazil | Seoul, South Korea     |
| ## | 60                   | 37                | 58                     |
| ## | Shanghai, China      | Shenzhen, China   | Singapore, Singapore   |
| ## | 38                   | 48                | 54                     |
| ## | Sydney, Australia    | Taipei, Taiwan    | Tehran, Iran           |
| ## | 43                   | 42                | 48                     |
| ## | Tokyo, Japan         | Vancouver, Canada | Washington DC, USA     |
| ## | 42                   | 54                | 54                     |

### 3.3 Detección de valores faltantes

|    |                                     |                         |
|----|-------------------------------------|-------------------------|
| ## | segment_id                          | city_name               |
| ## | 0                                   | 0                       |
| ## | admin_ward                          | latitude                |
| ## | 0                                   | 0                       |
| ## | longitude                           | catchment_id            |
| ## | 0                                   | 0                       |
| ## | elevation_m                         | dem_source              |
| ## | 161                                 | 0                       |
| ## | land_use                            | soil_group              |
| ## | 0                                   | 0                       |
| ## | drainage_density_km_per_km2         | storm_drain_proximity_m |
| ## | 284                                 | 239                     |
| ## | storm_drain_type                    | rainfall_source         |
| ## | 0                                   | 0                       |
| ## | historical_rainfall_intensity_mm_hr | return_period_years     |
| ## | 0                                   | 0                       |
| ## | risk_labels                         |                         |
| ## | 0                                   |                         |
| ## | segment_id                          | city_name               |
| ## | 0.000000                            | 0.000000                |
| ## | admin_ward                          | latitude                |
| ## | 0.000000                            | 0.000000                |
| ## | longitude                           | catchment_id            |
| ## | 0.000000                            | 0.000000                |
| ## | elevation_m                         | dem_source              |
| ## | 5.433682                            | 0.000000                |
| ## | land_use                            | soil_group              |
| ## | 0.000000                            | 0.000000                |
| ## | drainage_density_km_per_km2         | storm_drain_proximity_m |

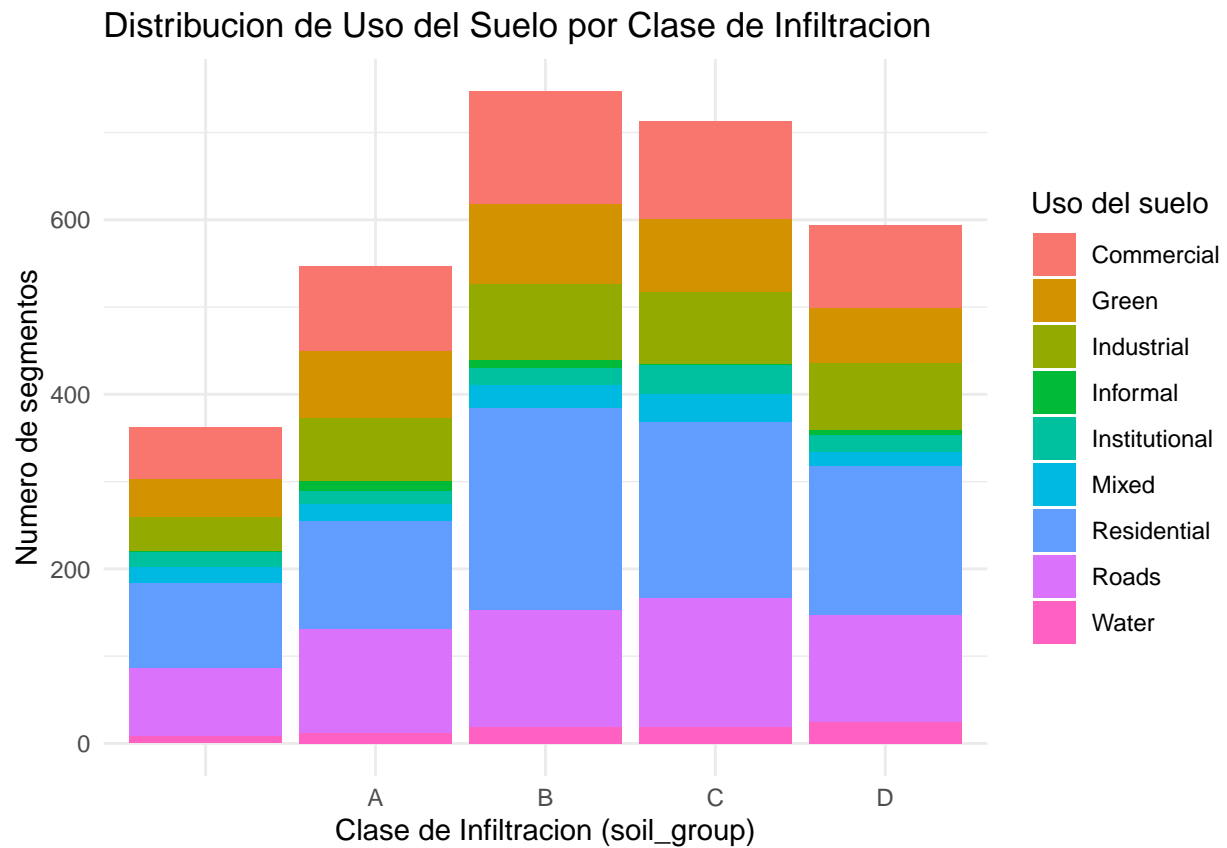
```
##          9.584880          8.066149
##          storm_drain_type          rainfall_source
##          0.000000          0.000000
## historical_rainfall_intensity_mm_hr          return_period_years
##          0.000000          0.000000
##          risk_labels
##          0.000000
```

### 3.4 Análisis adicional: Distribución de uso del suelo por grupo de suelo

```
## Cantidad de usos del suelo:
```

```
##
## Commercial      Green      Industrial      Informal Institutional
##          493          359          357          29          106
##          Mixed      Residential      Roads      Water
##          110          827          599          83
```

```
## Uso de suelo mas comun: Residential
## Frecuencia: 827
```



## 4. Revisión bibliográfica

### Estudios previos relacionados

- **UN-Habitat (2020)**: Destaca que las inundaciones urbanas son un riesgo creciente en ciudades de Asia y África debido a la urbanización y el cambio climático. Recomendando análisis de datos espaciales para identificar zonas vulnerables.
- **Jha et al. (2012)**: En “Cities and Flooding,” se analizan factores como densidad de drenaje y uso del suelo en inundaciones pluviales, usando modelos GIS y regresión.
- **Tingsanchali (2012)**: Estudio sobre Bangkok mostró que la baja elevación y alta impermeabilización aumentan el riesgo de inundaciones, usando modelos hidrológicos.

### Métodos usados en investigaciones similares

- **Análisis exploratorio**: Histogramas, boxplots y mapas de calor para correlaciones (usando R o Python).
- **Modelos predictivos**: Regresión logística, árboles de decisión y redes neuronales para predecir riesgos de inundación.
- **Análisis espacial**: Herramientas GIS (ArcGIS, QGIS) para mapear elevación y drenaje.
- **Técnicas de preprocesamiento**: Imputación de valores faltantes (media, mediana, KNN) y normalización de variables numéricas.

## 5. Plan de trabajo

### Metodología general

#### 1. Preprocesamiento:

- Limpieza: Imputar valores faltantes (ej. media para numéricas, moda para categóricas).
- Procesar `risk_labels` separando etiquetas múltiples.
- Verificar duplicados y valores atípicos.

#### 2. EDA avanzado:

- Visualizaciones: Mapas de calor (correlaciones), boxplots por ciudad, mapas con `latitude/longitude`.
- Correlaciones entre variables numéricas y `risk_labels`.

#### 3. Modelado:

- Clustering (K-means) para agrupar segmentos por riesgo.
- Regresión logística o árboles de decisión para predecir `risk_labels`.

#### 4. Validación: Dividir datos (70% entrenamiento, 30% prueba), evaluar métricas (precisión, F1-score).

### Cronograma de actividades

- Semana 1-2: EDA básico, limpieza inicial, revisión bibliográfica.
- Semana 3-4: Preprocesamiento avanzado, visualizaciones.
- Semana 5-6: Modelado inicial (clustering, regresión).
- Semana 7: Validación y redacción del informe final.
- Semana 8: Preparación de la sustentación.

## Herramientas y lenguajes

- **Lenguaje:** R (análisis, visualización, modelado).
- **Paquetes:** `dplyr`, `ggplot2`, `tidyr`, `caret` (modelado), `sf` (análisis espacial).
- **Herramientas:** RStudio, R Markdown para informes.

## 6. Expectativas y retos

### Posibles dificultades

- **Valores faltantes:** ~10% en `elevation_m`, `soil_group`, etc., pueden sesgar resultados.
- **Datos sintéticos:** Elevaciones negativas o distribuciones no realistas.
- **Complejidad de `risk_labels`:** Etiquetas múltiples requieren parsing (ej. separar `ponding_hotspot|low_lying`).
- **Desbalance:** Algunas ciudades o usos del suelo pueden estar sobrerrepresentados.

### Estrategias para resolverlas

- Imputar valores faltantes con media/mediana (numéricas) o moda (categóricas).
- Validar elevaciones negativas con análisis geográfico (ej. confirmar si son áreas costeras).
- Usar funciones de `tidyr` para separar `risk_labels`.
- Aplicar técnicas de balanceo (submuestreo o sobremuestreo) si hay desbalance.

## 7. Conclusión preliminar

El análisis inicial del dataset `urban_pluvial_flood_risk_dataset.csv` revela un problema crítico de inundaciones urbanas, con variables como `elevation_m`, `drainage_density_km_per_km2`, y `historical_rainfall_intensity_mm_hr` clave para identificar riesgos. El EDA muestra valores faltantes (~10% en algunas columnas) y distribuciones variadas en `land_use` (Residencial domina) y `soil_group`. Las hipótesis planteadas se explorarán con modelado predictivo. Los retos incluyen manejar datos faltantes y etiquetas múltiples, pero R ofrece herramientas robustas para avanzar. Este análisis sentará las bases para estrategias de mitigación en ciudades globales.

““